Many (though by no means all) of the questions that development economists and policymakers ask themselves are causal in nature: What would be the impact of adding computers in classrooms? What is the price elasticity of demand for preventive health products? Would increasing interest rates lead to an increase in default rates? Decades ago, the statistician Fisher proposed a method to answer such causal questions: Randomized Controlled Trials (RCT) (Fisher, 1925). In an RCT, the assignment of different units to...
different treatment groups is chosen randomly. This insures that no unobservable characteristics of the units is reflected in the assignment, and hence that any difference between treatment and control units reflects the impact of the treatment. While the idea is simple, the implementation in the field can be more involved, and it took some time before randomization was considered to be a practical tool for answering questions in social science research in general, and in development economics more specifically.

About twenty years ago, the idea of randomized controlled trials was just starting to make its way into development economics. Starting in 1994, Glewwe, Kremer, and Moulin (2009) kick-started the use of randomized evaluations among development economists and practitioners (Kremer, 2003). In 1997, the PROGRESA randomized controlled trial began, marking the first evaluation of a large scale policy effort in a developing country. With the launch of these randomized evaluations, we, perhaps naively, expressed the hope that randomized controlled trials would revolutionize social policy in the twenty first century, much as they had revolutionized medicine in the 20th century (Duflo and Kremer, 2003; Duflo, 2004; Banerjee et al, 2007). With the century less than 20 years old, it seems a little premature to evaluate this claim. Randomized evaluations clearly take a larger place in the policy conversation now than they did at the turn of the century and receive substantially more funding from donor organizations and local governments. Policy innovations that have been tested with RCTs have reached millions of people. However, the amount of money involved is still small. Development policy, moreover, is known for its twists and turns; many have predicted that RCTs are just the current fad and, soon enough, will have their comeuppance.

Something that we did not anticipate, however, has undoubtedly happened: randomized controlled trials have, if not revolutionized, at least profoundly altered, the practice of development economics as an academic discipline. Some scholars applaud this change (we are obviously in that camp), while others rue it (Ravallion, 2012; Deaton, 2010) but the fact is not really in dispute. In this essay, we start by quantitatively documenting this remarkable evolution. We go on to discuss the ways in which the field has been affected by the practice of randomized controlled trials, and what we see as their main contributions to the practice of development economics.
The popularity of RCTs as a research tool has sometimes been seen as conflicting with their potential (or ambition) for changing the world. The view is that the “academic” desire to come up with the cleverest research design may not line up with the practitioners need to identify scalable innovations (the next cell phone), or change “systems” (health care) or reform institutions (democracy). Using the USAID Development Innovation Ventures (DIV) portfolio as a case study, we identify the policy innovations tested with DIV funding that have eventually led to large scale reach (over 100,000 people). The analysis suggests that the proposed opposition between interesting and important is not particularly pertinent. In practice, many of the interventions supported by DIV that have reached this scale started as small research projects driven by academics. These projects also had the greatest “bang for the buck” evaluated in terms of life eventually reached per USAID initial funding dollars. We conclude this essay by discussing what this tells us about the policy process and the role RCTs can have in it.

1. Rapid growth

Over the last fifteen years, the use of experiments has expanded in academia and in international organizations: the DIME group at the World Bank lists over 200 studies, the vast majority of them randomized, and Arianna Legovini, the head of DIME, estimates that if we take the Bank as a whole, there are at least 475 randomized controlled studies going on (Legovini, personal communication). Tables 1 and 2 and Figures 1 to 5 summarize some of the trends in the use of experiments over time.

We start by a review of impact evaluations conducted by Cameron et al. (2015) (Figures 1 and 2). They compiled a repository of 2259 impact evaluation studies in development economics that were published between 1981 and 2012 by searching all major academic databases in health, economics, public policy and the social sciences. They supplemented this with an online crowdsourcing effort, which offered a $10 gift certificate per qualifying paper that was not already in the database. They then classified them by sector and by type. Overall, 66% (1491) of those evaluations are RCTs. Figure 2 shows that the number of RCTs has grown rapidly over time.

---

3 Which does not necessarily imply they have the highest social return.
Next, we look at the data compiled by Aidgrade (Vivalt, 2015). Aidgrade compiles the results of impact evaluations of development interventions. According to Vivalt:

The evaluations included in the AidGrade database were carefully selected from a number of different databases and online sources, the detailed process for which is outlined in Vivalt (2015). AidGrade.org employees first chose 30 topics they felt were important development issues. Those lists were combined and made into one large list of topics. The list was then narrowed down based on whether or not there were likely to be enough evaluations for a meta-analysis. The search universe includes search aggregators, such as Google Scholar and EBSCO, but also includes the J-PAL, IPA, CEGA, and 3ie only databases.

Figure 3 shows the number of evaluations per year, and Figure 4 shows how the evaluations ventilate over time between RCTs in economics, RCTs in other fields (e.g. medical trials) and non-RCTs. Both figures show a clear trend in both the number and the fraction of RCTs among the impact evaluations that are surveyed.

RCTs are particularly popular among younger researchers. Figures 4 and 5 show the number and the fraction of researchers who carry out RCT among the fellows and associates of BREAD, the association of development economists, by year in which they have obtained their PhD. The number clearly increases among the recent PhDs, and while this is in part driven by a larger number of recent fellows and associates, the fraction of them who do RCTs increases as well.

The number of RCTs presented at development economics conferences grew rapidly until 2010, and stabilized (or decreased) after that. At the annual conference of BREAD (the flagship conference in development economics), the fraction of papers featuring RCTs increased from 8% in 2005 to 63% in 2010, and hovered around 40-50% after that (except for the last conference, at Georgetown, where it was 28%). At NEUDC, a larger conference attended by many junior researchers, the fraction of RCTs has been fairly stable, ranging between 16% and 24% for the years 2012 to 2016 (the years for which we could get the papers) and showing no particular trend (Table 1).

RCTs have made a clear entry in top academic journals. Looking at AER, QJE, Econometrica, Review of Economic Studies and JPE, the number of RCT studies was 0 in
1990, 0 in 2000, and 10 in 2015 (Table 2). At the same time, the number of development papers published in these journals almost doubled (from 17 in 1990 to 32 in 2015). Table 2 also provides the details by journal. This is not driven by any particular journal (except that *Econometrica* does not seem to contribute much). Note that this does not mean that RCT studies have supplanted other type of work: the vast majority of published work on development is still non-RCT (if we take lower ranked journals), and even in top journals the experiments have been in addition to the (limited number of) papers that were published on development.

Beyond the growth in the number of experiments, and in the number of researchers who carry them out, what also stands out is the range and the ambition of the projects that are attempted: few topics seem off limit, and scale does not seem to be a barrier.

Researchers work directly with governments to randomize aspects of their work. Finnan, Olken and Pande (2015) describe several of these ambitious experiments. For example, Dal Bó, Rossi and Finan (2013) randomize the wages at which new government employees are hired, while Olken, Khan and Khwaja (2016) randomize incentives for tax collectors in Pakistan and Ashraf, Bandiera, & Lee (2015) work on how government health workers are recruited for their job. In experiments covering several districts and millions of workers Muralidharan, Niehaus and Sukhtankar (2016) and Banerjee et al. (2016) evaluate two separate process changes in the payment of wages of India’s major workfare program (MGNREGS), while Banerjee, Duflo and Keniston (2014) randomize reforms in the police department in India and Duflo, Greenstone, Pande, and Ryan (2013a and 2013b) randomize the enforcement of pollution regulation on industrial firms in India.

Researchers work at a scale that is sufficient to capture market equilibrium effects: Muralidharan and Sundararaman (2015) randomize a private school voucher at the school market level while Muralidharan, Niehaus and Sukhtankar (2016), in their aforementioned experiment, are able to look at the impact of MGNREGS on wages and productivity.

The range of topics keeps expanding. Development economists study alcohol addiction (Schilbach, 2015), electoral fraud in Afghanistan (Callen & Long, 2014), Cognitive Behavioral Therapy (CBT) for ex-combatants (Blattman et al., 2015), early childhood stimulation and development (Attanasio et al., 2014).
In summary, randomized experiments have become, not so much the “gold standard” as just a standard tool in the toolbox. Running an experiment is now sufficiently commonplace that by itself it does not guarantee that the paper would get into a top journal or even the BREAD conference. On the other hand, researchers from all sorts of perspectives have come to consider RCTs as a feasible option for answering the questions they are interested in. This level of comfort is in part due to the growth of several entities that help researchers with the fieldwork including by codifying and standardizing experimental practices, training enumerators, etc. The leader for this is Innovation for Poverty Action (IPA), with its vast network of country offices and experienced staff workers, but also J-PAL, CEGA, and the World Bank. There is also more funding available, from USAID (DIV in particular), the World Bank (SIEF and DIME), DFID, The Bill and Melinda Gates Foundation, The William and Flora Hewlett Foundation, The International Initiative for Impact Evaluation, in particular and, more recently, the Global Innovation Fund. But part of it also has to do with the appeal of the technique. In the next section, we reflect on the influence that RCTs have had on development economics research and why.

2. The influence of RCTs on development economics research
The remarkable growth in the number of RCTs, and more generally in the importance of empirical development economics as a field, is in itself a dramatic change. The type of development research that is carried out today is significantly different from research conducted even fifteen years ago. A reflection of this fact is that many researchers who were openly skeptical of RCTs, or simply belonged to an entirely different tradition within development economics (e.g. Daron Acemoglu, Derek Neal, Martin Ravallion, Mark Rosenzweig) are involved in one or more randomized controlled trials in a developing country.

Early discussions of the merits (or lack thereof) of randomization put a lot of emphasis on its role in the reliable identification of internally valid causal effects and the external validity of such estimates. We, and others, have had these discussions in other places (Heckman 1992; Banerjee, 2007; Duflo, Glennester, and Kremer, 2007; Banerjee and Duflo, 2009; Deaton, 2010), and we won’t reproduce them here. As we began to argue in Banerjee
and Duflo (2009), we actually think that these discussions somewhat miss the point about why RCTs are really valuable, and why they have become so popular with researchers.

a. **A greater focus on identification across the board**

The original motivation of randomized experiments, starting with Neyman (1923) (as a theoretical device) and Fisher (1925) (who was the first to propose physically randomizing units), was a focus on the credible identification of causal effects. As Athey and Imbens (2016) write in their chapter for *The Handbook on Field Experiments*:

> There is a long tradition viewing randomized experiments as the most credible of designs to obtain causal inferences. Freedman (2006) writes succinctly ‘experiments offer more reliable evidence on causation than observational studies.’ On the other hand, some researchers continue to be skeptical about the relative merits of randomized experiments. For example, Deaton (2010) argues that ‘evidence from randomized experiments has no special priority. …Randomized experiments cannot automatically trump other evidence, they do not occupy any special place in some hierarchy of evidence.’ Our views align with that of Freedman and others who view randomized experiments as playing a special role in causal inference. Whenever possible, a randomized experiment is unique in the control that the researcher has over the assignment mechanism, and by virtue of this control, selection bias in comparisons between treated and control units can be eliminated. That does not mean that randomized experiments can answer all causal questions. There are a number of reasons why randomized experiments may not be suitable to answer particular questions,

For a long time, observational studies and randomized studies progressed on largely parallel paths: in agricultural science, and then biomedical studies, randomized experiments were quickly accepted, and a vocabulary and statistical apparatus to think about them were developed. Despite the adoption of randomized studies in other fields, in the social sciences most researchers continued to reason exclusively in terms of observational data. The main
approach was to estimate associations, and then to try to assess the extent to which these associations reflect causality (or to explicitly give up on causality). Starting with Rubin’s (1974) fundamental contribution, researchers started to use the experimental analog to reason about observational data, and this set the stage for thinking about how to analyze observational data through the lens of the “ideal experiment”.

Through the 1980s and 1990s, motivated by this clear thinking about causal effects, labor economics and public finance were transformed by the introduction of new empirical methods for estimating causal effects (matching, instrumental variables, difference-in-differences and regression discontinuity designs). Development economics also embraced those methods starting in the 1990s, but unlike in labor economics and public finance, some researchers also decided that it may be possible to go straight to the “ideal” experiment or to go back and forth between experimental and non-experimental studies. This means that the two literatures developed in close relationship, constantly cross-fertilizing each other.

The non-experimental literature was completely transformed by the existence of this large RCT movement. When the “gold standard” is not just a twinkle in someone’s eyes, but the clear alternative to a particular empirical strategy and a benchmark for it, researchers feel compelled to think harder about identification strategies, and to be more inventive and rigorous about them. As a result, researchers have become increasingly more clever at identifying and using natural experiments, and at the same time, much more cautious in interpreting the results from them. Not surprisingly therefore, the standards of the non-experimental literature have improved tremendously over the last few decades, without necessarily sacrificing their ability to ask broad and important questions. To take some examples, Alesina, Giuliano and Nunn (2013) use suitability to the plow to study the long run determinants of the social attitudes towards the role of women; Padro-i-Miguel, Qian and Yao (2014) use a difference and difference strategy to study village democracy; and Banerjee and Iyer (2005) and Dell (2010) use a spatial discontinuity to look at the long run impact of extractive institutions. In each of these cases, the questions are approached with the same eye for careful identification as other more standard program evaluation questions.

Meanwhile, the RCT literature was also influenced by work done in the non-experimental literature. The understanding of the power (and limits) of instrumental
variables allowed researchers to move away from the basic experimental paradigm of the completely randomized experiment with perfect follow up and use more complicated strategies, including encouragement designs. Techniques developed in the non-experimental literature offered ways to handle situations in the field that are removed from the ideal setting of experiments (imperfect randomization, non-compliance, attrition, spillovers and contamination, etc.). Structural methods were combined with experiments to estimate counterfactual policies (Todd and Wolpin, 2006; Attanasio, Meghir and Santiago, 2011).

More recently, machine learning techniques have also been combined with experiments to model treatment effect heterogeneity (see Athey and Imbens (2016) for a recent review of the econometrics of experiments).

Of course, the broadening offered by these new techniques comes with the cost of making additional assumptions on top of the original experimental assignment, and those assumptions may or may not be valid. This means that the difference in the quality of identification between a very well-identified, non-experimental study and a randomized evaluation that ends up facing lots of constraints in the field or tries to estimate parameters beyond pure treatment effects is a matter of degree. In this sense, there has been a convergence across the empirical spectrum in terms of the quality of identification, though mostly because experiments have pulled the remaining study designs up with them.

b. Assessing external validity
In the words of Athey and Imbens (2016): “external validity is concerned with generalizing causal inferences, drawn for a particular population and setting, to others, where these alternative settings could involve different populations, different outcomes, or different contexts.” The question of the external validity of RCTs is even more hotly debated than that of their internal validity. This is perhaps because, unlike internal validity, there is no clear endpoint to the debate: heterogeneity in treatment effects across different types of individuals could always occur, or heterogeneity in the effect may result from ever-so-slightly different treatments. As Banerjee, Chassang and Snowberg (2016) acknowledge: “External policy advice is unavoidably subjective. This does not mean that it needs to be uninformed by experimental evidence, rather, judgment will unavoidably color it.”
It is worth noting that there is very little here that is specific about RCTs (Banerjee, 2009). The same problem afflicts all empirical analysis with the one exception of what Heckman (1992) calls the “randomization bias.” “Randomization bias” refers to the fact that experiments require the consent of both the subjects and the organization who is carrying out the program, and these people may be quite different. Glennerster (2016), in her chapter in the Handbook of Field Experiments, provides the list of the characteristics of the ideal partner, and they are clearly not representative of the typical NGO. On the other hand, it is worth pointing out that any naturally occurring policy that gets evaluated (i.e. not an RCT) is also selected: the evaluation requires that the policy did take place, and that was presumably because someone thought it was a good idea to try it out.

In general, any study takes place in a particular time and place, and that would affect results. This does not imply that subjective recommendations by experts, based both on their priors and the results of their experiments, should not be of some use for policymakers. Most policymakers are not stupid, and they do know how to combine the data that is presented to them with their own prior knowledge of their settings. From our experience, we have often observed that when presented with evidence from an RCT on a program of interest, the immediate reaction of a policymaker is to ask whether an RCT could be done in their own context.

There is one clear advantage that RCT's do offer for external validity, although it is not often discussed, and has not been systematically exploited as yet. To assess any external validity issues, it is helpful to have well-identified causal studies in multiple settings. These settings should vary in terms of the distribution of characteristics of the units, and possibly in terms of the specific nature of the treatments or the treatment rate, in order to assess the credibility of generalizing to other settings. With RCTs, because we can, in principle, control where and over what sample experiments take place (and not just how to allocate the treatment within a sample), we can therefore, get a handle on how treatment effects might vary by context. Of course, on its own, this is not sufficient to say anything much, if we account for infinite unstructured variation in the world. But there are several ways to make progress.
A first approach is to combine existing evaluations, and make assumptions about the possible distribution of treatment effects. Rubin (1981) proposes modeling treatment effect heterogeneity as stemming from a normal distribution: in each site, the causal effect of the treatment is a site-specific effect drawn from a normal distribution. The goal is to estimate the mean and the variance of the treatment effect, and the implied specific site effect, taking into account the fact that we have the other effects too. An interesting case study is the effect of microfinance programs. Meager (2016) analyzes data from seven randomized experiments, including six published in a special issue of the *American Economic Journal: Applied Economics* in 2015, and finds remarkable consistency in the mean effects across these studies, but much more heterogeneity in their variance. Of course, to carry out this exercise properly, we need access to an un-selected sample of studies, and since there is publication bias in economics, the sample of published studies may not be representative of all the studies that exist. This is where another advantage of RCT kicks in: since they have a defined beginning and end, they can in principle be registered. To this end, the American Economic Association recently created a registry of randomized trials (www.socialsciencesregistry.org), which, as of June 1, listed 699 studies. The hope is that all projects are registered, preferably before they are launched, and that results are clearly linked to the study, so that in the future meta-analysts can work from the full universe of studies.

A second approach is to conceive projects as multi-site projects from the start. One recent example of such an enterprise is the “Graduation” approach—an integrated, multi-faceted program with livelihood promotion at its core that aims to “graduate” individuals out of extreme poverty and onto a long-term, sustainable higher consumption path. BRAC, the world’s largest nongovernmental organization, has scaled-up this program in Bangladesh (Bandiera et al. 2013), while NGOs around the world have engaged in similar livelihood-based efforts. Six randomized trials were undertaken over the same time period across the world (Ethiopia, Ghana, Honduras, India, Pakistan, and Peru). The teams regularly communicated with each other and with BRAC to ensure that their local adaptations remain true to the original program. The results suggest that the integrated multi-faceted program was “sufficient” to increase long-term income, where long-term is defined as three years after the productive asset transfer (Banerjee et al., 2015). Using an index approach to
account for multiple hypotheses testing, positive impacts were found for consumption, income and revenue, asset wealth, food security, financial inclusion, physical health, mental health, labor supply, political involvement and women’s decision-making after two years. After a third year, the results remained the same in 8 out of 10 outcome categories. There is country-by-country variation (e.g. the program was ineffective in Honduras), and the team is currently working on a meta-analysis to quantify the level of heterogeneity.

One issue is that there is little that the researcher can do ex-post to reliably identify the source of differences in findings across countries. A third possible approach would be to take guidance from the first few sites to make a prediction on what the next sites would find. To discipline this process, researchers would be encouraged to use the results from existing trials to make some explicit predictions about what they expect to observe in other samples (or with slightly different treatments). These can serve as a guide for subsequent trials. This idea is discussed in Banerjee, Chassang and Snowberg (2016), who call it “structured speculation.” They propose the following broad guidelines for structured speculation:

1. Experimenters should systematically speculate about the external validity of their findings.
2. Such speculation should be clearly and cleanly separated from the rest of the paper, maybe in a section called “speculation”
3. Speculation should be precise, and falsifiable

Structured speculation has three advantages, according to Banerjee, Chassang and Snowberg (2016). First, it ensures that the researcher’s specific knowledge is captured. Second, it creates a clear sense of where else experiments should be run. Third, it creates incentives to design research that has greater externality. They write:

To address scalability, experimenters may structure local pilot studies for easy comparison with their main experiments. To identify the right sub-populations for generalizing to other environments, experimenters can identify ahead of time the characteristics of groups that can be generalized, and stratify on those. To extend the
results to populations with a different distribution of unobserved characteristics, experimenters may elicit the former using the selective trial techniques discussed in Chassang et al. (2012), and run the experiments separately for each of the groups so identified.

As this approach was just proposed recently, there are few examples as yet. A notable example is Dupas (2014). Dupas (2014) studies the effect of short-term subsidies on long-run adoption of new health products, and reports that short-term subsidies had a significant impact on the adoption of a more effective and comfortable class of bed nets. The paper then provides a clear discussion of external validity: It first spells out a simple and transparent argument relating the effectiveness of short-run subsidies to: 1) the speed at which various forms of uncertainty are resolved; 2) the timing of user’s costs and benefits. If the uncertainty over benefits is resolved quickly, short-run subsidies can have a long-term effect. If uncertainty over benefits is resolved slowly, and adoption costs are incurred early on, short-run subsidies are unlikely to have a long-term effect.

It then answers the question “For what types of health products and contexts would we expect the same results?” It does so by classifying potential technologies into three categories based on how short-run (or one-time) subsidies would change adoption patterns. Clearly, there could be such discussions at the end of all papers, not just ones featuring RCTs. But because RCTs can be purposefully designed and placed, there is a higher chance of follow-up in this case.

c. Observing the unobservable

If the main benefit of randomization is not the identification of causal effect, what is it? And what explains its remarkable success among researchers?

We agree with Athey and Imbens (2016) that “a randomized experiment is unique in the control that the researcher has over the assignment mechanism” and we would take the argument one step further: randomization is also unique in the control that the researcher (often) has on the treatment itself. In observational studies, however beautifully designed, the researcher is limited to evaluating what has been implemented in the world. In a
randomized experiment, she can manipulate the treatment in ways that we do not observe in reality. This has a number of advantages. First, she can innovate, i.e. design new policies or interventions that she thinks will be effective based on prior knowledge or theory, and test them even if no policymaker is thinking of putting them in practice yet. Development economists have many ideas, often inspired by what they have read or researched, and many of the randomized experiment projects come out of those: they test in the field an intervention that simply did not exist before (a kilogram of lentil for parents who vaccinate their kids; stickers to encourage riders to speak up against a bad driver; free chlorine dispensers, etc.).

Second, she can introduce variations that will help her establish facts that could not otherwise be established. The well-known Negative Income Tax (NIT) Experiment was designed with precisely that idea in mind: in general, when wages are raised, this creates both income and substitution effects which cannot easily be separated (Heckman, 1992). But randomized manipulation of the slope and the intercept of a wage schedule makes it possible to estimate both together. Interestingly, after the initial NIT and the Rand Health Insurance Experiment, the tradition of social experiments in the US, as Judy Gueron describes in her chapter in the Handbook of Field Experiments (Gueron, 2016), has mainly been to obtain causal effect of social policies that were often fairly comprehensive packages. In contrast, development economists have worked both on evaluations of real policies (e.g. the PROGRESA evaluation, or, more recently, the evaluation of the graduation program) but also on what Congdon, Kling, Ludwig, and Mullainathan (2016) describe as “mechanism experiments.” They write:

Broadly, a mechanism experiment is an experiment that tests a mechanism—that is, it tests not the effects of variation in policy parameters themselves, directly, but the effects of variation in an intermediate link in the causal chain that connects (or is hypothesized to connect) a policy to an outcome. That is, where there is a specified policy that has candidate mechanisms that affect an outcome of policy concern, the mechanism experiment tests one or more of those mechanisms. There can be one or more mechanisms that link the policy to the outcome, which could operate in
parallel (for example when there are multiple potential mediating channels through which a policy could change outcomes) or sequentially (if for example some mechanisms affect take-up or implementation fidelity). The central idea is that the mechanism experiment is intended to be informative about some policy but does not involve a test of that policy directly.

In other words, mechanism experiments do not confine themselves to testing feasible (or desirable) policies. For example, cars with broken windows could be put in the street to test the broken window theory. Once we realize that we are not limited to a set of realistic policy options (though we are constrained by what is ethically acceptable), this opens up a wide range of possibilities.

Banerjee and Duflo (2009) discuss some examples of mechanism experiments. One prominent example in development is a project conducted by Karlan and Zinman (2005) in collaboration with a South African lender that gives small loans to high-risk borrowers at high interest rates. The experiment was designed to test the relative weights of \textit{ex post} repayment burden (including moral hazard) and \textit{ex ante} adverse selection in loan default. Potential borrowers with the same observable risk are randomly offered a high or a low interest rate in an initial letter. Individuals then decide whether to borrow at the solicitation’s offer rate. Of those who apply at the higher rate, half are randomly offered a new, lower contract interest rate when they are actually given the loan, whereas the remaining half continue at the offer rate. Individuals did not know \textit{ex ante} that the contract rate could differ from the offer rate. The researchers then compared repayment performance of the loans in all three groups. The comparison of those who responded to the high-offer interest rate with those who responded to the low offer interest rate in the population that received the same low contract rate allows the identification of the adverse selection effect; comparing those who faced the same offer rate but differing contract rates identifies the repayment burden effect. The basic idea of varying prices \textit{ex post} and \textit{ex ante} to identify different parameters has since been replicated in several different studies (e.g. Ashraf, Berry and Shapiro (2010) and Cohen and Dupas (2010)). The experimental variation was key here, and not only to avoid
bias: In the world, we are unlikely to observe a large number of people who face different offer prices, but receive the same actual price.

Experiments can also be set up to understand the way institutions function. An example is Bertrand et al. (2007), who set up an experiment to understand the structure of corruption in the process of obtaining a driving license in Delhi. They recruited people who are aiming to get a driving license and set up three groups, one that receives a bonus for obtaining a driving license quickly, one that gets free driving lessons, and a control group. They found that those in the “bonus” group get their licenses faster, but those who get the free driving lessons do not. They also found that those in the bonus group are more likely to pay an agent to get the license (who, they conjecture, bribes someone). They also found that the applicants who hired an agent were less likely to have taken a driving test before getting a driving license. Although they did not appear to find that those in the bonus group who get licenses are systematically less likely to know how to drive than those in the control group (which would be the litmus test that corruption does result in an inefficient allocation of driving licenses), this experiment provides suggestive evidence that corruption in this case does more than “grease the wheels” of the system.

Such designs do not always directly lead to actionable policy, but they have allowed us to describe or understand how the world works. For example, in the seminal Bertrand and Mullainathan (2004) study, researchers sent resumes to prospective employers. The resumes are paired, such that there are identical resumes, except for the name of the job applicants, who can either be white sounding or African American sounding. They find that “applicants” with black sounding names are half as likely to be called back as those with white sounding names. Furthermore, being highly educated does not help, which suggests that something other than statistical discrimination is at play. This design has been replicated hundreds of times in different settings, providing extensive evidence of discrimination against different people and in different markets. This large body of evidence does not necessary point to a specific solution to this problem, or even helps us determine the root of this behavior, but, unlike the previous literature, it provides clear evidence that the phenomenon exists.
d. Data collection

Experiments have also spurred creativity in measurement. In principle, there is no automatic link between careful and innovative collection of microeconomic data and the experimental method. And, indeed, there is a long tradition in development economics to collect data that is specifically designed to test theories: both the breadth and the quantity of microeconomic data collected in development economics has exploded in recent decades, not only in the context of experiments (see Udry (1995) for a prominent early example).

However, one specific feature of experiments that serves to encourage the development of new measurement methods is high take-up rates and a specific measurement problem. In many experimental studies, a large fraction of those who are intended to be affected by the program are actually affected. This means that the number of units on which data needs to be collected to assess the impact of the program does not have to be very large and that data are typically collected especially for the purpose of the experiment. Elaborate and expensive measurement of outcomes is therefore easier to obtain than in the context of a large multipurpose household or firm survey. By contrast, observational studies must often rely on variation for identification (policy changes, market-induced variation, natural variation, supply shocks, etc.) that cover large populations, requiring the use of a large data set often not collected for a specific purpose. This makes it more difficult to fine-tune the measurement to the specific question at hand. Moreover, even if it is possible 

Some of the most exciting recent developments in empirical development economics have to do with measurement. Researchers have turned to other sub-fields of economics, as well as entirely different fields, to borrow tools for measuring outcomes. Examples include soil testing and remote sensing in agriculture (see de Janvry, Sadoulet and Suri (2016) for a review on agriculture); techniques developed by social psychologists for difficult to measure outcomes, such as audit and correspondence studies, implicit association tests, Goldberg experiments and List experiments (see Bertrand and Duflo (2016) for a review of their use to
measure discrimination); tools developed by cognitive psychologists for child development (Attanasio et al., 2014); tools inspired by economic theory, such as Becker-DeGroot-Marshak games to infer willingness to pay (see a discussion in Dupas and Miguel, 2016); biomarkers in health, beyond the traditional height, weight and hemoglobin (cortisol to measure stress for example); wearable devices to measure mobility or effort (Rao, Schilbach & Schofield, in progress; Kreindler, in progress).

Specific methods and devices that exactly suit the purpose at hand have also been developed for experiments. Olken (2007) is one example of the kind of data that can be collected in an experimental setting. The objective was to determine whether audits or community monitoring were effective ways to curb corruption in decentralized construction projects. Getting a reliable measure of actual levels of corruption was thus necessary. Olken focused on roads and had engineers dig holes in the road to measure the material used. He then compared that with the level of material reported to be used. The difference is a measure of how much of the material was stolen, or never purchased but invoiced, and thus an objective measure of corruption. Olken then demonstrated that this measure of “missing inputs” is affected by the threat of audits, but not, except in some circumstances, by encouraging greater attendance at community meetings. Rigol, Hussam & Regianni (in progress) provide another example of clever data collection methods. For their experiment, they designed soap dispensers that could track when the pump was being pushed in order to accurately measure if and when people wash their hands, and hired a Chinese company to manufacture the dispensers. Similar “audit” methodologies are used to measure the impact of interventions in health, such as patients posing with specific diseases to measure the impact of training (Banerjee et al., 2016) or ineligible people attempting to buy free bed nets (Dupas et al., 2016). Even a partial list of such examples would be very long.

In parallel, greater use is being made of administrative data, which are often combined with large-scale experiments. For example, Banerjee et al., (2016) make use of both publicly available administrative data on a workfare program in India and restricted expenditure data made available to them as part of the experiment; Khan, Khwaja and Olken (2016) use administrative tax data; and Attanasio, Medina and Meghir (2016) use unemployment insurance data to measure the long term effect of job training in Colombia.
The bottom line is that there has been great progress in our understanding of how to creatively and accurately collect or use existing data that go beyond the traditional survey, and these insights have led both to better projects and to innovations in data collection that have been adopted in non-randomized work as well.

c. **Iterate and build on previous research in the same settings**
The next methodological advantage of RCTs also relates to the control that researchers have over the assignment and, often enough, over the treatments themselves. Well-identified policy evaluations often leave us with many questions on why things turned out the way they did. For example, a number of papers using regression discontinuity designs find that the impact of “elite” schools on the marginal child who is admitted tends to be very low. These results seem to hold both in rich and in poor countries (Abdulkadiroglu, Angrist, and Pathak, 2014; Dobbie and Fryer, 2014; Lucas and Mbiti, 2014; Dustan, de Janvry, and Sadoulet, 2015; Clark, 2009). But this leaves a number of questions pending: does this mean that the impact is zero for all students or just the marginal student? Is it because peers don’t matter and curriculum doesn’t matter or because they both matter but cancel out?

While some progress can be made (for example, Abdulkadiroglu, Angrist & Pathak (2014) exploit the fact that students take two different tests to get a handle on the impact of the program for different types of students), one is necessarily limited by the type of policy variation that is actually available. The result from a single RCT often likewise raises more questions than it can actually answer. For example, when Duflo, Kremer and Robinson (2008a) found that the return to fertilizer appears to be very large even when used by the farmers themselves on their own fields (and not just on experimental plots), one possible policy response might have been to follow Jeff Sachs’ idea in distributing fertilizer for free. But this was not their next step: instead they started wondering why farmers are not using more fertilizer. This set them down a path that led them to set up a number of experiments in the same setting: some focused on learning and social networks, and some on the difficulty to save even over short periods of time. This latter inquiry led them down the path of designing and implementing a specific product where the household was offered the option of buying fertilizer in advance (Duflo, Kremer and Robinson, 2008b). The social
network interventions found surprisingly little diffusion of agricultural innovation to immediate friends, and this set them down another path: how could it be the case, given all we know about how much people talk about agriculture? To unpack this further, they introduced a simple device designed to address a problem that they noticed in their first set of experiments: households tend to overuse fertilizer (conditional on using it), relative to what appears to be profit-maximizing. They then set up a number of experiments to study in what conditions this device does spread, and what this tells us about how farmers decide whether to talk to each other and trust each other (Duflo, Kremer, Robinson, Schilbach, in progress).

Analyzing these results will no doubt spur new questions and experiments. All empirical science is of course iterative, with studies building on each other. But the ability to work in the same setting, with the same outcome and measurement, is extremely precious, and not available outside of a controlled setting.

f. Unpacking the interventions
Finally, RCTs, allow the possibility to “unpack” a program, to its constituent elements. Here again the work may be iterative. For example, all the initial evaluations of the BRAC ultra poor program were done using their “full package,” as were a large number of evaluations of the Mexican CCT program PROGRESA. But both for research and for policy, once we know that the full program works, there is a clear interest in knowing why it works. In recent years, a number of papers have looked “inside” CCT, relaxing the conditionality, for example. Some work has been conducted on the role and the type of conditionality, (see Baird, McIntosh and Ozler (2011), Bursztyn and Coffman (2012) and Benhassine et al. (2015) for examples), followed by many papers experimentally varying other features (we return to the impact of this work below).

Similarly, the early results of the evaluation of the ultra poor program have set the stage, both for a more theoretically grounded understanding of exactly which market failures led to a poverty trap, as well as a more practically grounded understanding of whether all of the interventions were truly necessary or if certain components could be removed. In the event that some components are unnecessary, costs could be lowered considerably, allowing the
program to reach more people using the same budget. Hanna and Karlan (2016) discuss how one could go from the initial “full package” evaluation to this greater understanding:

The ideal method, if unconstrained by budget and organizational constraints, is a complex experimental design that randomizes all permutations of each component. The productive asset transfer, if the only issue were a credit market failure, may have been sufficient to generate these results, and if no other component enabled an individual to accumulate sufficient capital to acquire the asset, the transfer alone may have been a necessary component. The savings component on the other hand may have been a substitute for the productive asset transfer, by lowering transaction costs to save and serving as a behavioral intervention which facilitated staying on task to accumulate savings. Clearly it is not realistic in one setting to test the necessity or sufficiency of each component, and interaction across components: Even if treated simplistically with each component either present or not, this would imply 2x2x2x2 = 16 experimental groups.

Several studies have tackled pieces of the puzzle, and more are underway (see the review in Hanna and Karlan, 2016). The way forward is clearly going to be the development of a mosaic, rather than any one definitive study that both tests each component and also includes sufficient contextual and market variations that it can help set policy for a myriad of countries and populations. More work is needed to tease apart the different components: asset transfer (addresses capital market failures), savings account (lowers savings transaction fee), information (addresses information failures), life-coaching (addresses behavioral constraints, and perhaps changes expectations and beliefs about possible return on investment), health services and information (addresses health market failures), consumption support (addresses nutrition-based poverty traps), etc. Furthermore, for several of these questions, there are key, open issues for how to address them; for example, life-coaching can take on an infinite number of manifestations. Some organizations conduct life-coaching through religion, others through interactive problem-solving, and others through psychotherapy approaches (Bolton et al., 2003; Bolton et al., 2007; Patel et al., 2010). Much
remains to be learned not just about the promise of such life-coaching components, but how
to make them work (if they work at all).

In some settings, particularly when working on a large-scale with a government, it is
actually possible to experiment from the beginning with various versions of a program. This
serves two purposes: it gives us a handle on the theory behind the program, and it has
operational value for the government, who can pick the most cost-effective combination.
Banerjee et al. (2015) is an example of this approach. The government of Indonesia was
interested in reducing corruption in their rice distribution program (Raskin), which is
infamous for reaching few of its intended beneficiaries and for not always been sold at the
right price. They thought that delivering a card to the beneficiaries with the eligibility
information might ameliorate this problem, and lead to greater benefits. Working with the
Government of Indonesia, the authors designed a set of field experiments to provide
information directly to eligible households. In 378 villages (randomly selected from among
572 villages spread over three provinces), the central government mailed “Raskin
identification cards” to eligible households to inform them of their eligibility and the
quantity of rice that they were entitled to. To unbundle the mechanisms through which
different forms of information may affect program outcomes, the government also
experimentally varied how the card program was run along three key dimensions— whether
an additional rule (the copay price) was also listed on the card, whether information about
the beneficiaries was also made very public, and whether cards were sent to all eligible
households or only to a subset. The researchers then collected data on eligible and ineligible
rice purchases and price paid for all villages. On net, they found that the card did lead to
large increases in the amount of subsidies received by the households. Further, they found
that the information on the card matters: the price paid was lower when the price was
indicated on the card. They also found that the card was more effective when the
information was made public. Finally, public information is not sufficient on its own, the
physical card also matters.

Knowing all of this is important to understand the mechanisms at play. It was also
immediately actionable for the government, who proceeded to scale-up the program, and to
provide cards with price information, to all eligible households accompanied by posters.
Cards were distributed to over 65 million individuals. This is one occasion where the researchers’ and the government’s interests were exactly aligned. Is it more generally true?

3. **Have RCTs become too academic to lead to any real world changes?**

RCTs have changed development economics but have they also had significant influence in the world? If RCTs are pushing forward the frontiers of academic research by seeking to understand mechanisms and testing ideas generated by academics themselves, does this make them too academic and less useful for policy?

In this section we argue that RCTs can contribute to policy not only by providing evidence on specific programs that can be scaled, but also by changing the general climate of thinking around an issue. We then examine a case study of a funder, Development Innovations Ventures at USAID. Some of the innovations that it has funded were driven by social entrepreneurs without researcher involvement and some were tested using RCTs and/or had close involvement of development economics researchers. A review of this portfolio suggests that several programs involving development economics researchers and RCTs had substantial real-world influence.

a. **Are RCTs that are more “academic” less useful for policy?**

Many studies seek not to test just a particular program, but also to contribute to a body of literature that seeks to test different theories of human behavior. If citizens vote for candidates based on their ethnicity or caste is that because of very strong preferences, clientelistic networks, or a combinations of weak preferences and no alternative information on candidate quality? Do people only value what they pay for? How important are liquidity constraints, as opposed to lack of information or low human capital, in explaining poor child health and low business profitability in low-income families?

The studies that seek to answer these questions do not always test standard development programs, although some time they may turn into development ideas. De Mel, McKenzie, and Woodruff (2012) gave cash to businesses in Sri Lanka without conditions, repayment requirements, or mentoring, something unheard of in finance programs at the time (of
course, eventually, the idea of unconditional cash transfers caught on as a realistic policy option, as indicated by the success of GiveDirectly). As we have discussed above, a series of studies that focused on pricing of health goods first asked households if they were willing to purchase a good at one price and then gave them the good at a lower price or for free, not something a regular program would do. Researchers pushed to test unconditional cash transfers (Baird et al., 2011; Blattman et al., 2014; Haushofer and Shapiro, 2013; Benhassine et al., 2014), even though at the time the political consensus was on conditional transfers.

The reason why this is potentially important for policy, and not just for academic curiosity, is that even where certain program specifics do not generalize, underlying patterns in human behavior may. The finding that small incentives are effective in encouraging people to take actions that have short-run costs but long-run benefits is more likely to generalize than the finding that lentils are a successful incentive for vaccination in Rajasthan (Banerjee and Duflo, 2010). Kremer and Glennerster (2011) review over 70 health economics RCTs and find strong similarities in consumer behavior across countries and products, including sharp reductions in take-up of non-acute care health products with small increases in price, big increases in take-up of non-acute products with small incentives (negative prices), and no evidence that paying for something makes people more likely to use it (Kremer and Miguel, 2007; Cohen and Dupas, 2010; Ashraf, Berry, and Shapiro, 2010a; Dupas, 2013).

This body of work on prices was taken up by advocates of free distribution of Insecticide Treated Bednets (ITNs). For many years there had been a fierce debate on the merits of free distribution, with free distribution advocates arguing that even small prices deter the poor, while others argued that small copayments were important to ensure ITNs were utilized. Armed with the evidence from RCTs, advocates of mass free distribution have successfully pushed this approach resulting in a dramatic rise in ITN coverage across Africa from roughly 2009 to 2015. The WHO reports that 43 of 47 countries in sub-Saharan Africa with ITN distribution programs provide them for free (World Malaria Report 2015). A recent article in Nature (Bhatt et al., 2015) examines the sharp decline in malaria infections in sub-Saharan Africa and estimates that between 2000 and 2015 malaria interventions prevented 663 million malaria cases, most of which is attributable to the sharp rise in ITN
coverage: 450 million cases of malaria prevented by ITNs, and roughly 4 million deaths from 2000 to 2015.

Beyond the specific example of malaria, the policy community is coming to a more general realization that higher prices for preventive health products can sharply decrease take up and that price elasticity of demand can be very high (Kremer and Holla, 2009; Kremer and Glennerster, 2011; Dupas, 2016). This is changing the entire approach to pricing of these products.

Another area where a body of evidence from RCTs has produced both specific policy changes and given rise to more general lessons that have profoundly changed the policy debate is on attitudes toward cash transfer programs. Arguably the biggest innovation in anti-poverty and social protection policies in developing countries over the past twenty years is the growth of Conditional Cash Transfer programs (CCTs). Beginning in Mexico, these have now spread to more than thirty countries, and they have arguably played an important role in the decline in poverty in Latin America (Alzua, Crues, and Ripani, 2013; Attanasio et al., 2005; Barrera-Osorio et al., 2011; Galiani and McEwan, 2013). While many factors were at play in the spread of Conditional Cash Transfers, we, along with many others, think the PROGRESA experiment (Schultz, 2004; Gertler, 2004) and the many following experiments in other contexts4 played a significant role, through influencing Mexico’s decision to continue and expand CCTs after the inauguration of a new administration, the active promotion of CCTs by the Inter-American Development Bank and the World Bank, and the adoption of CCT by many countries.

More recently, there has been additional examination of how conditional cash transfers work that is further changing the policy debate. Conditional cash transfers have been shown by RCTs to not only increase the behavior on which the cash is conditional but to also improve outcomes such as height, weight, and cognitive development (Barham et al., 2013) and reduce HIV infection (Baird et al., 2011). There is also no evidence that poor households spend increased cash on alcohol or other temptation goods (Haushofer and Shapiro, 2013; Masterson and Lehmann, 2014; Evans and Popova, 2014). Indeed, the

---

evidence suggests that the income elasticity of demand for food out of cash transfers is surprisingly high (see a review in Banerjee (2015)), and food transfers do not improve nutrition more than cash transfers (Cunha, 2014).

This is leading to a movement from a situation in which policymakers would almost never consider cash transfers to one in which cash transfers, conditional or not, are becoming an accepted tool in development policy. For example, as the world struggles to cope with refugees from war, groups such as the International Rescue Committee have drawn on RCTs of cash distributions in stable environments and with refugees (Masterson and Lehmann, 2014) to strongly push for cash rather than in-kind support for refugees. David Miliband, IRC president and CEO, said:

The spate of man-made and natural disasters enveloping innocent civilians raises profound questions not just for international politics, but for NGOs and the humanitarian sector, as well. If we keep doing ‘business as usual,’ the gap between need and provision will continue to grow. Cash distribution – alongside clear humanitarian ‘floor’ targets in the revised Millennium Development Goals, more sustainable local partnerships and better use of evidence overall – could be part of a vital renewal of the humanitarian sector.

Early in the introduction of RCTs, Lant Pritchett (Pritchett, 2002) argued that RCTs would never become particularly popular with policymakers because they have reason to prefer ignorance over rigorous knowledge, in order to continued favoring their preferred program: “It pays to be ignorant.” While in some cases policymakers may have incentives to preserve ignorance, in other cases policymakers are aware of the holes in their knowledge and would like to learn more. They may have a strong attachment to a favorite program, either due to inertia or a political imperative. But the experience of running the program often persuades them that they could do it better, and they are surprisingly open to ideas about how to improve their programs. The Raskin and MGNREGS programs mentioned above, where several teams of researchers have worked with the government, are good examples—while it was clear that the programs would continue, finding ways to make them work better was clearly of interest.
b. How to assess the policy success (or not) of the RCT agenda

It is a little bit difficult to assess the causal effect of RCTs on policy adoption. Interventions subject to RCTs are not themselves randomized, and many factors influence whether and when a particular intervention is taken up. When a program is taken up after an RCT showed it has worked, it is not always because of the RCT, and it is never just because of the RCT. Nevertheless, some have argued that the influence of RCTs on policy is actually quite low, compared to the volume of RCTs. For example, Shah, Wang, Nadel and Fraker (2015) point out that despite the 489 completed J-PAL evaluations, there are only 9 scale-up or policy influence stories on J-PAL’s web site. But this number per se is not particularly informative: for one, it is not a census of the studies that have some impact. Not all J-PAL RCTs are systematically followed up. These stories are chosen precisely because of the size of their impact and because they can be documented clearly. The absolute number of lives reached by them is quite significant—the J-PAL website tells us that over 200 million people were reached by these programs. But the main concerns with any statistic like this are conceptual:

1. The J-PAL web-site does not carry statistics on non-J-PAL studies for the very good reason that, based on our experience collecting information from DIV and J-PAL, it is far from straightforward to collect information on the extent to which RCTs have influenced policy. This means, for example, that the number does not include the hundreds of millions of people who have been reached by CCT.
2. Many RCTs are fairly recent. Taking these to the policy level requires a lot of care, especially given the external validity issues (Would it work in government? Would it work in a different place?) The process is therefore often slow, again for very good reasons. Therefore, we should not expect a lot of these to be scaled as yet.
3. Many of the most valuable RCTs are those that test popular and highly touted policies that already exist in the world on a large-scale and show that they are in fact much less effective than previously claimed or believed. Microfinance and improved cook-stoves are two obvious examples. In such cases, success would be to slow down the spread of such policies. Obviously, in such cases, one would not expect
something to appear on the JPAL scale-up page, but these are two cases where the work has probably been quite influential.

4. In some cases, the primary purpose of an RCT is not to directly affect policy, but to instead investigate an underlying theoretical mechanism, which may, in turn, indirectly influence policy. However, such cases would not appear on a list of scale-ups, despite the fact that the knowledge they have provided has impacted, albeit indirectly, a large number of people. For example, the orthodoxy in development economics had long been that the poor are “poor but rational.” The accumulating evidence from RCTs has undoubtedly hastened the diffusion of the idea into development economics and into development policy that poor people are not always rational: this is reflected for example, in both the content, and the number of RCTs in the World Development Report (2016) on psychology and poverty. In turn, publications like the WDR, and the associated discussions, influence the design of policies.

5. It is not clear what the right benchmark for success should be. We suspect that if one looked at other areas of economics, one would find that research projects influenced policy at a much lower rate than RCTs have in development policy in recent years. Moreover, one would not want to say that rapid policy influence is the sole or even the major metric by which the worth of economic research should be assessed—think of the idea of congestion pricing for road use (Vickrey, 1969), which is only beginning to find real world applications.

6. Perhaps most importantly, it is worth realizing that the payoff to RCTs is likely to be the average of a highly skewed distribution. Looking at the fraction of RCTs which scale, rather than the average payoff, is therefore as misleading as looking at the fraction of any research and development effort which succeeds in terms of say generating a successful marketed product, since the payoff to research and development in general is typically very highly skewed. As is well-known, citations across scientific disciplines appear to follow a power law distribution, with a small fraction of papers accounting for the majority of citations. This peak is followed by a steep decay, as a large portion of research papers are never cited (Radicchi,
As we mentioned, the 9 policy innovations listed by J-PAL together have reached over 200 million people, and this does not include the more than 100 million people who have been reached through India’s most recent round of deworming, the millions of people who have received free bed nets (since J-PAL lists it as policy influence but does not provide a count), and the 60 million people whose water and air is less polluted because of the statewide adoption of better regulation of industrial pollution in Gujarat (again, not counted).

7. For this reason, pointing out that many research and development efforts yield low payoffs does not suggest that these are bad investments ex ante. The correct analytical question to ask is whether the expected average or marginal payoff to R&D effort in RCTs is positive or greater than that in other areas of research if one takes overall research budgets as fixed. Of course, measuring the payoff to research is inherently a difficult exercise for all sorts of conceptual reasons. There is also the added statistical difficulty that a large amount of data is needed to accurately measure the mean of a fat-tailed distribution.

c. **What have we learnt from the DIV experience?**

Keeping all of this in mind, we now turn to one particular example, the experience of the investments made by USAID’s Development Innovation Ventures (DIV) between 2010 and 2012.

DIV holds a year-round grant competition for innovative solutions to a range of development challenges, pilots and tests them using analytical methods, and scales solutions that demonstrate widespread impact and cost-effectiveness. DIV supports novel business or organizational models; operational, behavioral or production processes; and products or services that can help address development challenges. DIV’s tiered-funding model, provides

---

5 For instance, in the social sciences in general, papers receive on average 0.5 citations in the first two years of publication, including self-citations (Klammer and Dalen, 2002), whereas in mathematics, medicine, and education the number is estimated to be less than 1 (Mansilla et al. 2007). Given the skewed distribution, this implies that the median paper is never cited. Similarly, the majority of new patents have extremely low value with a small fraction of patents accounting for much of the overall value of patents.
small grants to pilot innovations in development; medium-size grants to rigorously test for impact and cost-effectiveness (often using RCTs) or ability to pass a market test; and larger-scale grants to help transition innovations to scale that have passed a market test or that have rigorous evidence of impact and cost-effectiveness.

When DIV was established, two targets were set for the program: 1) a 15% social rate of return on investment, and 2) a reach of at least 75 million people worldwide, through direct investment and through broader influence on the rest of USAID. Preliminary work by DIV staff suggests that the 2010-2 portfolio easily met the first goal, even under the conservative assumptions that all innovations supported by DIV yielded no further benefits, and even looking only at a subset of innovations that yielded financial benefits or health benefits that could be valued in terms of DALYs. While social return is a more conceptually comprehensive measure for evaluating DIV, it is difficult to measure, and this piece seeks not to evaluate DIV, but rather to look at the narrower question of whether RCTs can have real world influence. We therefore focus on examining the number of people reached by innovations supported by DIV (as well as by later adapted versions of these innovations). (Note that substantial reach is a necessary but not sufficient condition for high social return, since the total social benefit of an innovation equals the net benefit per person reached times the number of people reached.) This exercise is inherently limited, so readers will have to make their own judgements about the likely impact per person reached, the likely future reach of these innovations (sustainability), and the extent to which DIV funding played an important role in the reach achieved by innovations in the DIV portfolio. What we are doing here is rather the descriptive exercise of systematically tracking a portfolio.

Nevertheless, following the entire 2010-2012 DIV portfolio is interesting for a paper that explores the influence of RCTs, because the premise of DIV is specifically to fund innovations in development that have the potential to cost-effectively reach a large number of people through either the public or the private sector.

In particular, whereas many other programs have a top-down approach in which program staff identify problems in advance, choose sectors on which to focus, or set strategy within sectors, DIV follows a bottom-up approach that is deliberately open across sectors: supporting innovations that will scale commercially, innovations designed to scale through
the public sector, and startups and organizations proposing to change behavior within existing large organizations. Although the bulk of DIV’s outreach effort has been oriented towards traditional social entrepreneurs, DIV has also made an effort to be open to proposals from development economics researchers. To balance this openness, DIV employs a staged finance approach in which innovations only receive larger-scale support after they have passed rigorous tests. DIV provides large-scale support (Stage 3) only for innovations which have rigorous evidence of impact and cost-effectiveness or which have demonstrated market viability. At the piloting (Stage 1) and testing stages (Stage 2), however, DIV has historically been open to proposals that have the potential to scale based on their cost-effectiveness, for example, even if they do not necessarily already have a management team in place capable of scaling internally or written commitments from scaling partners.6

This combination of approaches thus helps us ask whether the engagement with the development economics research community, and the willingness to consider early-stage investments even without a fully proven capacity to scale, came at the cost of scaling success. We can shed light on these questions by comparing the scaling record across types of projects, stages of funding, and of course by looking at the scaling record of DIV.

In the Appendices, we provide a list of all the DIV awards from this time period, and a narrative description of the innovations that have, subsequent to DIV’s funding, reached more than 100,000 people.

Table 3 shows the results of this exercise. Here are some key insights:

1.) **DIV has been relatively successful in supporting innovations that scale.** A relatively high fraction of DIV awards, and an even higher fraction of DIV total investment, went to projects which, to date, have already reached more than 100,000 people (and a smaller but still high fraction of the awards went to projects that reached more than a million people).

---

6 Although DIV does not require a proven pathway to scale at Stages 1 or 2, a promising pathway to scale through the public or private sector (or a hybrid of the two) and strong potential demand is one of its main selection criteria, particularly at Stage 2.
30% of DIV awards (13/43) have so far reached more than 100,000 people within 3-5 years. These awards account for 57% of the total value of DIV awards in this time period, or $10.98 million in total funding. 14% of DIV awards (6/43) have so far reached more than one million people. These awards account for 33% of the total value of DIV awards in this time period, or $6.38 million in total funding.

Why do we say that 30% is “relatively successful”? A rule of thumb in the venture capital world is that 10% of investments yield modest success and 1% yield large successes. While we have not yet identified other funders that publish data that would allow for computation of comparable statistics, our reading of the literature and our examination of websites of some other organizations suggests that these rates compare well with those achieved over a much longer time frame by other impact-investing organizations. These results are all the more striking because, while some organizations provide funding only after a certain level of scale is reached (e.g., Acumen, Skoll Foundation), DIV often supported innovations at an early stage (as well as tests to know whether they were worth scaling up), rather than waiting until innovations had already reached a certain scale and had attracted earlier support before investing.

2.) **Stage One and Stage Two awards have a particularly low DIV expenditure per person reached and account for more than 90% of people reached by innovations supported by DIV during this time period.**

One of these early stage innovations (Consumer Action and Matatu Safety) recently received a Stage 3 DIV award, but in general, Stage 1 and 2 innovations attained high levels of reach because other funders/entities provided support based in part on the information generated from the DIV-funded project.

3.) **While the estimated DIV expenditure per person is lower for earlier stage grants, it is fairly low across the board. This is because the great majority of**

---

Footnote: 7 Two innovations (that reached over 100,000 people) received both a Stage 1 and a Stage 2 award. Thus, these 12 awards support 10 separate innovations.
the reach of DIV-supported innovations was attained without the applicants returning to DIV for additional financial support.

Though many past-awardees apply for additional funding, only 7% of DIV’s 2010-2012 portfolio of grantees received follow-on funding after the initial period of performance. Over 40% of DIV’s 2010-2012 grantees received follow-on funding from either the public or private sector after DIV’s investment. DIV’s capacity to be catalytic of course partly derives from the rich funding ecosystem in which it operates, where other entities (governments, NGO, private sector firms) can adopt innovations.

4.) **Cost was a key determinant of which innovations scaled. The largest scale was achieved by innovations with very low costs per person.**

In some cases, the innovations involved the provision of information by media or phone (e.g., voter report cards, election monitoring), or provided behavioral “nudges” in large, existing systems (e.g., Zambian community health workers). Of course it’s important to recognize that total impact depends on the benefit per person reached times the number of people reached, and some innovations with moderate cost per person (e.g. VisionSpring) and moderate reach may generate high total social benefit because the benefit per person is very high.

5.) **While some innovations reached more than 100,000, or in one case, more than 1,000,000 people through the creation and growth of a new organization designed to scale the innovation, the vast majority of reach was delivered through adoption by existing large organizations, including large firms, NGOs, and governments.**

Four of the DIV-supported innovations which reached 100,000 or more consumers involved the creation of new organizations which scaled from scratch. Seven involved adoption of the innovation by existing entities that already had high levels of reach.
Of the six innovations which reached more than one million people, one was scaled by an NGO which constructed and built operations around the innovation (Evidence Action in the case of chlorine dispensers), and four did so by adoption by existing organizations (an insurance company and the Kenyan National Transport and Safety Authority in the case of stickers in matatus, the Government of India in the case of biometric monitoring, political campaigns in the case of real-time efforts to send polling station outcomes to central locations by mobile phones, and newspapers in the case of voter report cards). Existing organizations with large reach that adopted DIV-supported innovations or modified versions of these innovations included private sector firms, NGOs, and governments.

6.) **Innovations tested with RCTs scale not only through adoption by governments, but also through adoption by private sector firms and NGOs.**

Of the ten DIV awards for innovations with RCTs that have reached more than 100,000 people, there were two clear cases in which developing country governments played the lead role (i.e. scaling of an improved approach to community health worker recruitment by the government of Zambia and biometric monitoring in India). The Kenyan government seems likely to play an important role alongside the insurance industry in scaling the Kenyan matatu safety program. Donors played a key role in provision of Potential Energy’s improved cookstoves in Darfur. NGO partners played a role in a number of projects. A major lesson of this analysis is that large private firms played a major role as well (e.g., an insurance company played a key role in the matatu stickers project and newspapers published the free content when an NGO provided them with voter reports cards).

7.) **Innovations involving randomized controlled trials, and/or developed in part by researchers (often working in close conjunction with implementers), reach 100,000 or 1,000,000 users at a particularly high rate.**
43% (10/23)\(^8\) of awards for which an RCT was used for evaluation or development economics researchers were involved in design of the innovation reached more than 100,000 people.\(^9\) 26% (6/23) of these awards supported innovations that had reached more than one million people in the original or adapted form (e.g. voter report cards, election monitoring, stickers in matatus, chlorine dispensers, and biometric attendance verification). In contrast, among the innovations not including an RCT component or a strong role for development economics researchers, only 16% (3/19) reached 100,000 people (Vision Spring, Mera Gao, d.light), and none reached more than one million people.\(^{10}\)

One could imagine multiple hypotheses for this difference in the rates of success. First, it might be easier to reach many people by persuading large organizations and governments to adopt the innovation and in this process the evidence from the RCTs might have played an important role. By contrast those innovations that did not come from the academic RCT side tried to scale by directly implementing or selling their product, which may be harder, as these innovations do not have large pre-existing policies, programs or institutions as initial partners. Second, it is often argued that academic researchers mainly want to publish, and this conflicts with their incentives to get involved in projects that are socially useful but not as creative (replication, tinkering with design, etc.). But on the other hand, it is also argued that journals have a strong publication bias, and it is easier to publish things that have worked. Ergo, development economists should have strong incentives to develop and test innovations that have a reasonable chance of success. Moreover, perhaps (just perhaps?)

---

\(^8\) Projects were coded as having development economics researchers involved if the initial proposal that was funded by DIV explicitly included the efforts of researchers. Although d.light’s initial proposal included an RCT on the impacts of their products, this RCT did not take place and funding strictly supported the development of a new solar home system as well as an \textit{ex post} impact evaluation of these systems. Due to these circumstances, we have not included d.light in our calculation of projects developed in part by researchers in this point. If we were to include d.light, this figure would be 11/24, or 46%.

\(^9\) Voter information report cards (2 awards), election monitoring technology, digital attendance and medical information systems in primary health care centers, mobile tools for community health care workers (2 awards), consumer action on Matatu safety, bringing safe water to scale, improved cookstoves, and recruiting community health workers.

\(^{10}\) 24 awards incorporated an RCT component or were based on an RCT. This excludes two cases in which the initial proposal included an RCT but the ultimate actual project funded by DIV did not include an RCT: Psychometric Analysis for Entrepreneurs (AID-OAA-F-13-00028) and Affordable Access to Energy for All: Innovative Financing for Solar Systems (AID-OAA-F-13-00007). Note that since there is a lot of overlap between researcher-led projects and projects with an RCT, we cannot easily separate their impact.
economics actually gives them some useful insights into the design of projects. Third, it may also be that the recent focus on information and behavioral economics makes them particularly interested in innovations with a low cost per user (“nudges”), which seems to be a strong predictor of success. Fourth, when researchers were involved, they were typically not just evaluators: they were fully involved in the development of the innovation (e.g. voter report cards, chlorine dispensers, monitoring project in Afghanistan), worked closely with implementing organizations, and remained closely involved in the details of the implementation. They were in fact “researcher-entrepreneurs”. Many of the ideas developed by researchers drew on the latest ideas in the field, and the data suggest that the researchers who developed these ideas were then relatively successful in working with others to scale these innovations.

8.) **Innovations that had already been tested through RCTs and found to have impact and potential for cost-effectiveness prior to applying for DIV support accounted for three of the five innovations that reached over one million people.**

Three of the five innovations that reached over one million people (voter report cards, Consumer Action and Matatu Safety, and Chlorine Dispensers for Safe Water) had already been subject to RCTs before applications were submitted to DIV. While we have not yet coded the data, we believe that there were very few applications in this category, so the rate at which proposals in this category reached over one million people was very high (possibly 100%).

9.) **While some DIV-supported innovations have been applied in multiple countries, most have not.**

So far, DIV-supported innovations have typically not been applied much beyond the country where they have been tested. This may be an area where future work is needed.
Conclusion

The discussion in this policy section suggests that RCTs have influenced policy both through providing evidence on individual projects and programs, and by changing thinking in development more broadly.

The biotech and IT industries routinely build on innovations developed by researchers using frontier techniques in those fields. The evidence from DIV awards is consistent with the idea that a similar approach may be effective in development, with innovations developed in part by researchers and/or involving randomized controlled trials reaching 100,000 or 1,000,000 users at a particularly high rate. This is absolutely not to say that work is not needed to fine tune interventions for different contexts, or that it is not important to evaluate real world programs that have not yet been evaluated using an RCT; but the development of new ideas that are grounded in basic science actually can lead to real-life change.

One striking lesson of this analysis is that the projects that are scaled up tend to be low-cost, well-defined, and simple. Other examples, not in this list, also fit this bill (e.g., deworming, the Raskin card). There are notable counterexamples of programs that are neither particularly cheap nor simple and have scaled up: Conditional Cash Transfers and the BRAC ultra poor programs are two examples. Furthermore, those two programs were not only scaled up where they had been tested, but were implemented in many other countries as well. Interestingly, they were initially replicated as RCTs.

Well-defined interventions are also the ones that are more likely to lead to successful research projects, since they can more easily pin down a specific mechanism, and be construed as a test for a theory. So the reasons why RCTs have been so successful as a research tool may also be what makes them successful at leading to real world changes.

Looking forward, we don’t know what the most important pathways of influence for RCTs might turn out to be. One route is that simple, clear insights, low-cost interventions, or low-cost modification to promising existing programs get adopted, as the DIV case study suggests. The fact that these innovations are low-cost of course does not mean that they have low impact. One lesson from decades of well-identified development research is that
details are incredibly important, and that the distinction between “big” and “small” questions can be very misleading (see Banerjee and Duflo (2012), Chapter 10 for a longer discussion).

An alternative pathway is one in which more complex interventions are replicated in many contexts and then widely adopted, following the PROGRESA or the BRAC model. The third one is that rather than just focusing only on the results, policymakers and other actors adopt the experimental attitude, i.e. leave some space within their operations for innovations and learning perhaps housed inside a specialized unit (like the White House “nudge” unit) or a cross-department fund (like the Tamil Nadu innovation fund).

But to really get the full benefits of the RCT revolution, it is not enough to do more RCTs and get some of them scaled up. A range of complementary institutions are also necessary to more effectively translate research into policy. For example, we need better systems for the production of meta-analyses and review articles and for the creation of expert panels to review the evidence. Medicine has a quite involved system for this, but even setting aside the question of how well that system works in medicine (Sim, 2001; Kawamoto et al., 2005), the institutions that are appropriate for medicine are not necessarily those that are appropriate for social science, and development economics in particular. These institutions are just starting to get built: the AEA registry of RCTs is an example of a successful effort to build a registration platform. Its popularity suggests that the development community is receptive to these efforts.

In addition to the purely scientific infrastructure for learning, the process of going from an idea to a program at scale requires appropriate institutional support. First, funders are needed to finance iterative piloting before an RCT to work out the implementation details. Once an RCT has been conducted, institutional support is also needed for iterating on the intervention to “ready” it to transition to scale. This includes testing ways to: bring unit costs down (since the first RCT often evaluates a small pilot with high unit costs); collaborate with potential implementing partners; and mitigate potential cost increases and/or reduced benefits that may results from institutional and personnel differences between the pilot and scaled-up versions of an innovation (due to, for example, government procurement systems.

---

11 Development Innovation Ventures and the Global Innovation Fund – a private fund modeled after DIV and to which DIV and other bilateral donors and impact investors contribute – explicitly encompass such a piloting phase.
with higher transaction costs or limited government capacity to implement the intervention effectively). To get to the right scaled-up version therefore involves trying them out to scale and measuring the impact at scale. Indeed, multiple iterations may be needed until something that is appropriate for policy can work. Figuring out how best to do the scaling in each case or how to do so in additional countries takes time, specialized human capital and additional funding.
References:


Attanasio, Orazio, Camila Fernández Emla O A Fitzsimons, Sally M Grantham-McGregor, Costas Meghir, Marta Rubio-Codina, Using the infrastructure of a conditional cash transfer program to deliver a scalable integrated early child development program in Colombia: cluster randomized controlled trial BMJ 2014;349:g5785


Banerjee, Abhijit, Jishnu Das and Reshmaan Hussam. 2016. "Improving the Quality of Private Sector Health Care in West Bengal” Forthcoming.


Callen, Michael, and James D. Long. "Institutional corruption and election fraud: Evidence


Duflo, Esther “Scaling up and Evaluation” (2004) ABCDE Annual World Bank Conference on development economics,


McKenzie D., S. de Mel, and C. Woodruff (February 2012). “One-time transfers of cash or capital have long-lasting effects on microenterprises,” Science, 335(24): 962-66.


Anxiety Disorders in Primary Care in Goa, India (MANAS): A Cluster Randomised Controlled Trial.” *The Lancet* 376 (9758): 2086–95.


APA


Figures & Tables:

**Figure 1: Number of Published RCTs**

**Figure 2: Evaluations by Type**
Figure 4. Fraction of BREAD Affiliates & Fellows with 1 or more RCTs

* Total Number of Fellows and Affiliates is 166.

Figure 5. Percent of BREAD Concurrence Papers using a RCT
<table>
<thead>
<tr>
<th>Journal</th>
<th>Year</th>
<th>Total # of Papers</th>
<th># of Development Papers</th>
<th># of which are RCTs</th>
</tr>
</thead>
<tbody>
<tr>
<td>AER</td>
<td>2015</td>
<td>101</td>
<td>15</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>2000</td>
<td>48</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>1990</td>
<td>57</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>QJE</td>
<td>2015</td>
<td>40</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>2000</td>
<td>43</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>1990</td>
<td>52</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>JPE</td>
<td>2015</td>
<td>36</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>2000</td>
<td>51</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>1990</td>
<td>65</td>
<td>9</td>
<td>0</td>
</tr>
<tr>
<td>Restud</td>
<td>2015</td>
<td>48</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>2000</td>
<td>36</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>1990</td>
<td>40</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Econometrica</td>
<td>2015</td>
<td>46</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>2000</td>
<td>37</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>1990</td>
<td>64</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>2015</td>
<td>271</td>
<td>32</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>2000</td>
<td>215</td>
<td>21</td>
<td>0</td>
</tr>
<tr>
<td>Year</td>
<td>Total # of RCTs</td>
<td>Percent RCTs</td>
<td></td>
<td></td>
</tr>
<tr>
<td>------</td>
<td>----------------</td>
<td>--------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2015</td>
<td>40</td>
<td>18.2%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2014</td>
<td>36</td>
<td>17.9%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2013</td>
<td>49</td>
<td>24.3%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2012</td>
<td>27</td>
<td>16.0%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 3: Future reach of DIV projects, by award type

<table>
<thead>
<tr>
<th>Award Stage</th>
<th>Number of Awards</th>
<th>Total Awarded Value</th>
<th>Fraction Reaching More than 100,000 people</th>
<th>Fraction Reaching More than 1,000,000 people</th>
<th>People Reached&lt;sup&gt;12&lt;/sup&gt;</th>
<th>DIV Expenditure per Person Reached</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stage 1</td>
<td>23</td>
<td>$2,353,136</td>
<td>17% (4/24)</td>
<td>8% (2/24)</td>
<td>6,723,733</td>
<td>$0.35</td>
</tr>
<tr>
<td>(&lt; $100,000)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stage 2</td>
<td>19</td>
<td>$9,557,926</td>
<td>44% (8/18)</td>
<td>11% (2/18)</td>
<td>16,931,044</td>
<td>$0.56</td>
</tr>
<tr>
<td>(&lt;$1,000,000)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stage 3</td>
<td>1</td>
<td>$5,516,606</td>
<td>100% (1/1)</td>
<td>100% (1/1)</td>
<td>1,750,000</td>
<td>$3.15</td>
</tr>
<tr>
<td>(&lt;$15M)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<sup>12</sup> Two innovations (Voter Information Report Cards and CommCare) that reached over 100,000 people received both a Stage 1 and a Stage 2 award. In both of these cases, people reached by those innovations are counted as people reached by Stage 2 awards.