Field Experiments and the Practice of Economics [1]

Prize Lecture, December 8, 2019 by
Abhijit Vinayak Banerjee
Massachusetts Institute of Technology (MIT), Cambridge, MA, USA.

WHEN, SOME TWENTY-FIVE YEARS AGO, I first started doing randomized controlled trials (RCTs), the most common reaction was one of puzzled tolerance. Colleagues and friends seem to admire the effort involved and could see that RCTs had the advantage of avoiding the then common wrangling about what is causal and what is not. But in the end they were skeptical that it was worth it. In part they were concerned for me – I had a successful career doing economics (in particular economic theory) the way it was done then. Why go down this particular rabbit-hole? But more importantly, as the more candid among them put it, "are RCTs economics?"

The prize in economics has perhaps settled that question at one level. But the doubts that motivated it remain. Responding to them provides a useful way to bring out what RCTs have brought to the field. At the risk of some caricature, there are really four closely related questions:

i. Economics aspires to generate *generalizable* knowledge. RCTs focus on estimating the impacts of *specific* interventions. Aren't these fundamentally different ways of approaching the world?

---

1. Esther Duflo and I gave and wrote up two lectures with parallel titles. They are companion papers and should probably be read together.

ii. Economics tackles *big* questions. RCTs by their very nature provide narrow and specific answers. How do we square that gap?
iii. Economics is about cumulatively building a theory, by building on the existing theories and making use of any new pieces of evidence to enrich the theory. Aren't RCTs piecemeal – one insight here, one insight there?
iv. Why are economists needed to run RCTs? Wouldn't it be better to have competent applied statisticians or the World Bank run them?

In this lecture I will make the case that each of these questions is based on a misunderstanding of both what economics today actually does and how knowledge from RCTs is used. This is partly because economics has changed – in part as a result of what Angrist and Pischke (2010) call the credibility revolution. And in part because RCTs have evolved from their initial adherence to the model set up by medical trials. I will use a number of examples to make these points. While they each highlight different issues, they each contribute something in response to all four questions.

2. GENERALIZABLE KNOWLEDGE

All science aspires to generalizable knowledge. If there was no hope of generalizing from what we learn in a specific study, it would be worth very little and researchers would not take it on. This is true even in fields like anthropology that emphasize the specificity of a particular moment in space and time. The point there is often to fight what anthropologists view as irresponsible or false generalizations.

2.1 Generalizing results from RCTs

RCTs come with a very simple approach to generalization. The idea is to implement the same concept in multiple locations to build confidence in its impact (or to discover that it only works in certain circumstances). In other words, the approach to generalization is to a substantial extent statistical, combining the results from multiple studies using statistical models that properly weight them (based on their precision etc.). The underlying assumption is that the impact is in part common across many locations (perhaps only those that are not too dissimilar) and over time, though there may also be an idiosyncratic component specific to each location/time. Meager (2019) is a nice example of this kind of statistical aggregation and shows, for example, that the impact of microcredit on the earnings of the average beneficiary was very similar across widely different locations and in fact, at best, quite small.

It is important however to emphasize the fact that it still is a concept that is being generalized and not an intervention. Every implementation is at least slightly different – the implementers might speak a different language or dress in a different way, for example. By treating each implementation as a different version of the same concept we are clearly imposing our own theory of what exactly matters. That theory could be wrong – we may have missed a key detail – but that is the chance we take. On the flip side, this makes clear that the process of looking for generalizable ideas is also a process of refining the theory of what exactly makes the intervention have impact (or not). For that reason, the set of treatments often evolves as locations get added, as we both pare down the original intervention to make it closer to the concept (perhaps while keeping the original as another intervention, at least to start with) and add other interventions that build on the insights from what has already been learnt. This process is described in some detail, for the case of primary education, in Banerjee *et al.* (2016) and in Esther Duflo's concurrent Prize lecture (2020). I will describe another example towards the end of this essay.

It is worth adding that this process of evolving interventions distinguishes RCTs from the often very similar research strategies that are bunched under the term "natural experiments". With natural experiments the basic idea of generalization is usually the same as with RCTs, but the extent to which multiple interventions are similar or different is often not in our control.

2.2 Generalization in conventional economics

Unlike RCTs, where the solution to the problem of generalizability is to generate new data, conventional economics treats the data we have as given. Something happened in a particular location at a particular time that changed the economic possibilities faced by different people and we are lucky enough to have data that tells us what changed in their lives. Typically, however, we would not know everything else that might have also been changing in their economic environment, and therefore how to identify the true causes of change. RCTs offer an advantage here – because of random assignment to treatment and control, we can be reasonably confident that any observed difference between the two populations can be causally ascribed to the treatment. However, that is widely accepted and has little to do with the issue of generalizing from the evidence.

There are two routes to generalizing policy-relevant ideas in conventional economics. The first is to suggest that behavior is likely to be universal because it is "consistent with economic theory". Economic theorists build models, which are toy universes where they deliberately assume away much of the complexity that we experience in everyday life, in order to be able to highlight specific mechanisms that might operate in the world. For example, in the standard model of labor supply, the key assumption is that the psychological cost of having to work a certain number of hours is independent of how much the workers are paid, their general standard of living, or their perceived well-being. Under this assumption, giving people a transfer they have not earned reduces the amount people work, which is often seen as a warning against wanton generosity.

The problem, of course, is that the highlighted assumption about the cost of having to work is not obviously true. For example, receiving a monthly cash gift from the government or an NGO might make the recipient less stressed about how they will make ends meet and therefore perhaps more productive. This is the idea proposed by Mullainathan and Shafir (2013). We will return to this and other related ideas in some detail, but a useful starting point is that being standard does not by itself make an assumption true.

However, in most cases the fact the theory predicts a particular behavior is not enough of a guide to policy. This is where our second route comes in. Even if what we called the key assumption were true, the choice of the actual policy, say whether or not to make an unearned transfer of X pesos a month, depends critically on the size of the labor supply response to the transfer. If the transfer reduces labor supply by a small enough amount, it might still be well worth making, since beneficiary families will end up richer on net. So, we do need to know the size of the labor supply response.

Trying to answer this question is when the enterprise becomes much more challenging. The standard protocol is to estimate a model of labor supply decisions using whatever data we have on how much people earn from various sources, and how hard they work. The parameters of the estimated model are then used to tell us the potential impact of specific interventions.

Estimating the model typically requires assuming something quite specific about the nature of the utility function for consumption (how fast the benefit of an extra dollar falls off at different levels of consumption) and a shape for the function representing the psychological cost of working (how much worse an extra hour of work is at different levels of work effort). For simplicity, we usually severely limit how much these costs and benefits vary across people and also what other features of their circumstances influence these functions. In particular, it is common to make what above we called the key assumption. We also make assumptions about the exact shape of the cost of effort, but equally importantly, very specific assumptions about how the cost depends for example on the nature of the work (collecting trash or sitting in an office?), the environment, physical and social, in the workplace (Is it hot? Is it friendly?), and the home environment of the workers (say whether or not they live near their parents and can therefore rely on them for childcare). One might imagine that accounting for these features can be very important before confidently generalizing evidence from the place where, say, we had the data, to other places with somewhat different circumstances. Unfortunately, more often than not, researchers estimating models choose to ignore most of these complexities (or to think of them as un-modeled sources of variation in behavior that are, rather implausibly, unrelated to everything else that is going on in the model).

In addition, to infer preferences from people's observed choices we need to make assumptions about what individuals observe about the world, what they believe, and what goals they are pursuing with their choices. Typically, the assumption is that decision-makers are quite well-informed and sophisticated about the circumstances and implications of their choices, and that they do not deviate systematically from the choices predicted by the maximization of the assumed preferences under their actual constraints. We know from the large and growing literature in behavioral economics that systematic errors are common in gathering and interpreting information, as well as in the choice process.

These rather stringent assumptions, as Todd and Wolpin (2010) acknowledge in their excellent review of this literature, come out of the basic dilemma of all empirical work. There is an inherent conflict between the number of parameters you can estimate (expanding the set of things that influence utility or cost, or allowing people to make systematic errors, potentially adds a very large number of parameters) and the precision with which you can estimate them. As a result, most attempts to recover the so-called deep determinants of behavior end up making a whole range of assumptions about the same behavior, many of which are untested and largely untestable given the data we have.[2]

Of course, this does not necessarily imply that the predictions of these kinds of models are always unreliable. Todd and Wolpin (2006) attempt to test the accuracy of model predictions by making use of the famous Progresa RCT in Mexico.[3] Progresa was a conditional cash transfer, which made payments conditional on children attending schools and receiving health check-ups. Todd and Wolpin estimate a model where families choose between sending their children to school and sending them to

2. There are of course many strategies to raise the plausibility of the model results. For example, the models often have multiple implications that were not used to estimate the parameters or data that was kept aside for testing the results.
3. Attansio, Meghir and Santiago (2012) do a similar exercise where, instead, they use both the treatment and the control group to estimate the model.

work, using just the data from the control group. They then compare the effect of the transfer estimated from that model to the estimate from the actual RCT. The results are a mixed bag – the model is quite accurate for girls but over-predicts the response to the subsidy for boys by a factor of two or three. And in many ways, this was a sort of best-case scenario for the prediction exercise – because of randomization, the control population was identical to the treatment population in terms of life and work environments. As a result, we avoided all the uncertainties that come in when we try to predict what would happen in a setting where, say, work choices are quite different.

2.3 The two ways to generalization

The advantage of the conventional approach is that it can be extremely economical – in principle, one data set can be used to estimate the many parameters needed to answer a whole range of policy questions – a much wider set than a typical RCT permits. For example, the Panel Study of Income Dynamics (PSID), which collects rich panel data on employment and income of individuals in the United States, has been used to estimate numerous parameters of relevance in labor economics, including income elasticities and risk preferences. This also means that we get to compare a very large set of policies. For instance, one could imagine studying how an income support program compares to a child-care subsidy as well as a road building effort, if the model being estimated was rich enough. In practice this does not happen much, again probably because of data constraints.

One obvious disadvantage is that the many assumptions that go into such an exercise often stretch credibility. Moreover, even after swallowing all the necessary assumptions, estimates of the various parameters do not always inspire confidence.[4] To take the example of a frequently used parameter, estimates of the widely-used coefficient of relative risk aversion vary between something like 0.2 and 10 or more, depending on the data and the methodology used to estimate it (Gandelman & Hernandez-Murillo 2014) and the important elasticity of labor supply with respect to the wage has a similar issue, though in recent years the best evidence seems to be converging towards a narrower range. Basing policies on these estimates clearly has its risks.

I see no reason, however, to restrict ourselves to one of these meth-

4. There is, for example, the issue that once one narrows down the parameters and has excess degrees of freedom, the choices of the moments that get used in the identification seem to make a big difference. This may well be for very good reasons – some moments may be better measured than others – but it does a layer of arbitrariness to the whole exercise.

ods. Especially given that RCTs have become so much easier to do, partly as a result of growing experience among researchers and wider acceptance in the policy community, and partly because of institutions like J-PAL and Innovations for Poverty Action (IPA), that there seems no reason not to make use them wherever possible. Indeed, it often clearly makes sense to combine them with the existing approaches of model estimation. It is now well-understood that results from RCTs are often ideal from the point of view of estimating a model both because of the richness of the data and because of better identification.[5] Conversely, the idea that we could first estimate a model to help us think about interventions to include in the RCT is an intriguing possibility. Even if the model estimates are not entirely reliable, they may provide useful bounds on what one could hope to achieve with specific interventions. Andreoni *et al.* (2019) is an interesting recent example, where the authors first use a model to estimate preferences for different groups of potential experimental subjects before assigning them to different incentive treatments.

The constraint at this point is that we do not have an effective enough language for talking about the relative reliability of different insights from these empirical exercises. This creates a tendency to either ignore everything except RCT results (including standard errors on the point estimates), to treat all results as deserving of equal weight, or, perhaps the worst, to pick and choose based on what fits the story. I must confess that I have a tendency towards the first kind of bias, though in Banerjee and Duflo (2011, 2019) we do make use of a wide range of evidence, including a lot of purely descriptive material. Improving our capacity to combine different forms of evidence in a more mindful way is an important next step for economics.

Generalizations are always partly an act of faith. However, this is no more so for policy conclusions coming from RCTs than for those coming from conventional economic policy analysis – if anything, less, because failures are easier to detect.

3. BIG QUESTIONS

One of the standard criticisms of RCTs is that they don't help us answer big questions: where is China headed? Is higher inequality necessary for faster growth? What kind of market economy is best?

More than a decade ago I wrote a piece titled, with a nod to Stephen Hawking, "Big Answers for Big Questions", arguing that while it is true

5. See for example, Attanasio, Meghir and Santiago (2012), Duflo, Hanna and Ryan (2012) and Banerjee, Duflo, Keniston, & Singh (2019).

that RCTs do not answer most of these kinds of "big questions", most other methods do not either, except by assertion or by ignoring the many frailties of the answers they offer. In particular, cross-country comparisons that often purport to answer these kinds of questions tend to be grossly unreliable, to the point of being nearly worthless for policy purposes. For example, on the question of the relationship between increases in inequality and increments in growth, Banerjee and Duflo (2003) show that a lot of the reported cross-country results ignore what the data actually says.

To the extent that the big questions can actually be reliably answered, RCTs may well be quite useful. For example, take the question of how to best design a market economy. China is surely the most striking economic success of recent years and, yet it is in many ways a very unusual market economy where the state owns a majority of the capital and controls the banking sector. The commonsense economics of 1989 would certainly not predict China to be a success, let alone the kind of success it has been, and in fact the Wall Street Journal asserted exactly this (i.e., China's purported failure) in its 100th anniversary issue from 1989. What we don't know is whether China would have been better or worse were it a more conventional market economy – and therefore whether Vietnam should emulate China or move in the opposite direction. Here the data cannot really help us, since there is no second China to compare with the first. On the other hand, one could imagine an RCT that helps us better design the institutions that underpin a good market economy. For example, pollution needs to be appropriately penalized for markets to work well, and the design of optimal incentives for pollution auditors was the subject of an RCT by Duflo *et al.* (2013).

3.1 What are big questions?

Perhaps more importantly, the definition of big questions is itself the product of a particular understanding of economics. The implicit and sometimes explicit premise is that the macroeconomy is key; in a market economy individuals are supposed to do the best they can within the constraints imposed by macroeconomic policies and the tax system. And yet, the evidence from many years of work in development economics suggests that this is not the case; markets routinely fail to deliver efficient outcomes and so do non-market institutions, like schools and hospitals run by governments and NGOs (Banerjee and Duflo 2011 summarizes this evidence). For a development economist the big questions are often whether people are realizing their full potential and, if not, what would enable them to do so.

The Bangladesh Rural Advancement Committee (BRAC), a remarkable

NGO, started a program in the early 2000s to help the poorest of the poor in the world's poorer countries "graduate" to a more normal poverty. They called it the Graduation Approach. The original program typically targets families that even other poor families consider very poor – often families that live off begging. The premise of the program is that even these households can become self-sustaining if given some initial help. The program offers these households a gift of productive assets (livestock, some working to start a vending business, etc.) of their choice, some temporary income support till their asset starts yielding a return, and a good dose of training, encouragement and hand-holding. The intervention typically ends after 18 months.

With funding from the Ford Foundation, 3ie, and U.S. Agency for International Development (USAID), J-PAL and IPA set up a consortium of researchers who conducted RCTs of this program in 7 countries – Bangladesh, Ethiopia, Ghana, Honduras, India, Pakistan and Peru. The results 18 months and 36 months following the ceasing of all external interventions show that treated households are substantially richer, healthier and happier than those not included in the program in all but one of the countries.[6] Moreover data from Bangladesh and India from both seven and ten years after the intervention started shows that consumption impacts continue to be as large as at 36 months, while the impacts on income actually grow. In other words, the intervention put these households on a new path, and they have never looked back.

This, to a development economist, is addressing a very big question – are those in extreme poverty there because they are intrinsically unproductive, or are they are just unlucky and caught in a poverty trap? The fact that program households are better off after ten years very much suggests the latter.

Why should such a trap exist? The basic idea is that the poor lack enough assets to be productive but the fact that they are not productive also makes them unable to accumulate the necessary assets. Underlying this hypothesis is the idea that financial markets are imperfect, which prevents the poor from borrowing the necessary assets. The rich stay rich for the same reason – they have enough assets to earn a high enough return on their assets to continue to be rich. There is a long tradition of papers that make this or related points, including some of the early work I did with Andy Newman (Leibenstein 1957, Dasgupta & Ray 1986, Banerjee & Newman 1993, Galor & Zeira 1993).

6. That country is Honduras, where the asset of choice was poultry and all the poultry died in an epidemic.

3.2 Are poverty traps real?

A beautiful recent paper by Balboni *et al.* (2019) provides more detailed evidence supporting the poverty trap story. They make use of the fact that in order for there to be a poverty trap, the relationship between past assets and current assets must be relatively flat at low levels of assets – the poor get stuck because they do not accumulate fast enough – but quite steep at some higher level, which allows the not-too-poor to escape extreme poverty. Balboni *et al.* explain this difference using transition diagrams (the relation between current wealth and past wealth) for the case where there is a poverty trap and one where there isn't. These are reproduced below

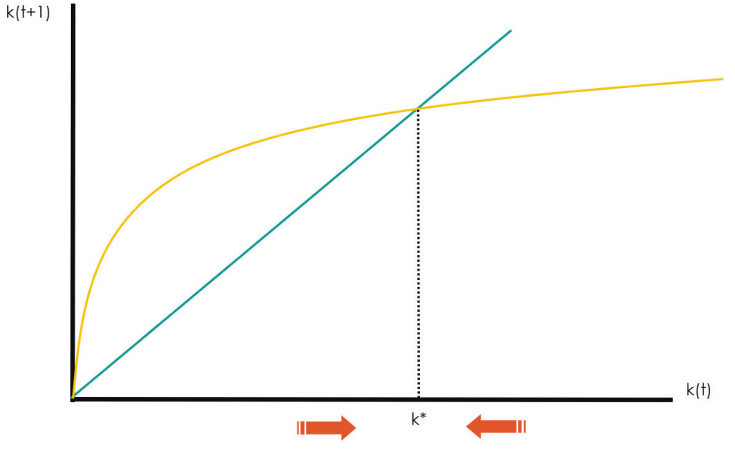

Figure 1. No Poverty Trap Case

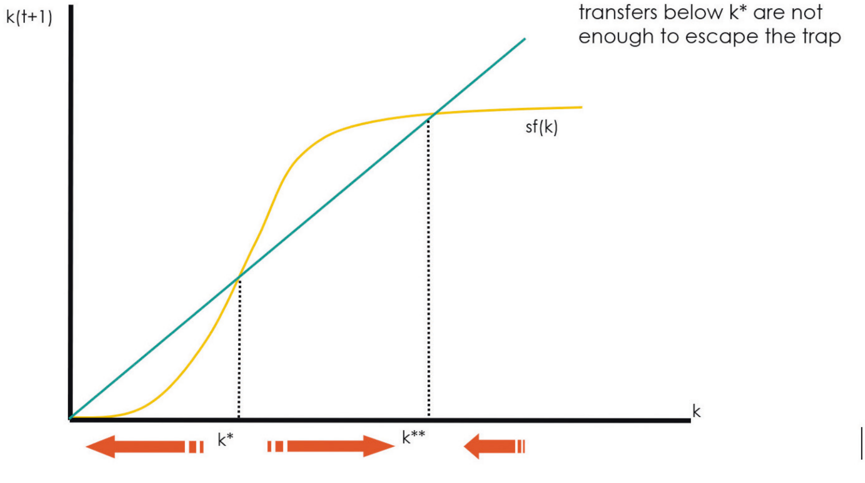

Figure 2. The Poverty Trap Case.

The authors plot the analogous relationship between past and current wealth in their data, taken from their study of the Bangladesh Graduation program mentioned above. As shown below, this empirical relationship definitely looks more like the second case than the first. A formal test confirms this as being the case.

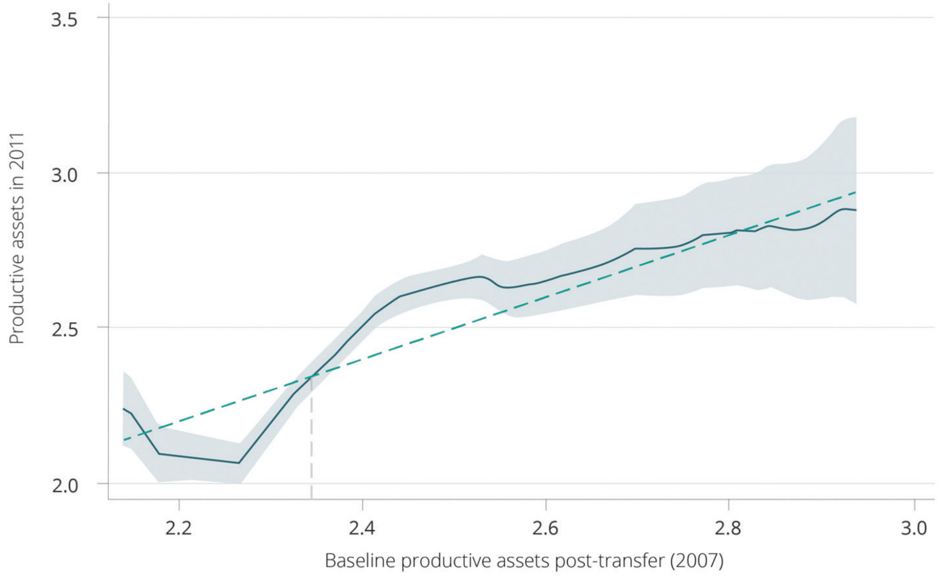

Figure 3. The empirical relationship.

One obvious implication of a transition diagram like the figure is that wealth paths should diverge – those who start just below the cutoff level (2.34 in the figure) will get poorer while those above get richer. This would mean that there should be relatively few households just around the cutoff. Instead, they should be clustered either significantly below the cutoff or significantly above. This is exactly what Balboni et al. find.

The fact that poverty traps are real is both good and bad news. Good news because it may be possible to liberate people from a life of extreme poverty with one push (like the Graduation program). The fact that income stays up on its own or even continues to grow following such transfers means that the cost-benefit ratio of such a program can be very favorable. Banerjee *et al.* (2015) report cost-benefit ratios that are well above 1.5 in most cases and as high as 4 in India. It is potentially bad news because uninsured shocks can throw households into a situation from which they cannot escape.

Either way it is a big deal.

4. THEORY

As already discussed in section 2, there is a tight relationship between theory, empirical work, and policy research in traditional economics. Interpreting evidence requires a model that rests on a body of theory. At the same time, estimation of the model informs our understanding of the theory because it tells us the parameter values needed to fit the data. Those parameter estimates should then influence future rounds of model building.

With RCTs (and natural experiments) the relationship is quite different. Policy implications are often direct – we do not necessarily have to go through the route of estimating a model. That means it may be possible to test even the most basic assumptions behind most standard models. Consider for example the assumption, highlighted in section 2, that the psychological cost of having to work a certain number of hours is independent of how much workers are paid, their general standard of living, or their perceived well-being. This is, as is well-known, at the heart of the concern that public generosity might lead to private sloth – the so-called income effect on labor supply is negative under this assumption. On the other hand, evidence from a set of RCTs across the world where households were randomly chosen to get a cash transfer with no work requirement shows the opposite. It is summarized by Banerjee *et al.* (2017) in a bluntly titled paper called "Debunking the Stereotype of the Lazy Welfare Recipient". Getting richer does not make poor people lazier.

4.1 Theory to experiments

Why does getting richer not make people lazier? One possibility is that our measurement of labor supply is so imprecise that we cannot detect any changes. Another is that the key assumption is incorrect. The traditional economic approach might be to try to estimate a model where the cost of work is allowed to depend on receipt of the cash transfer. The problem is that we do not directly observe the cost of effort – all we see is that program beneficiaries are not working less, despite being richer by the amount of the transfer. How do we know that it is not the benefits of effort rather than the cost that changed? Perhaps the extra money allowed some households to pay for childcare and therefore freed them up to work more hours. We have known since Benjamin (1992) that, in the presence of labor and credit market frictions, cash transfers can increase labor supply even if the standard model is correct, because they change the household's income earning possibilities. We therefore need further assumptions about how each of these effects plays out to separate the traditional income effect from this liquidity effect in model estimation.

What an RCT offers in this context is an ability to explore the source of the impact of cash transfers in much more detail. This is exactly what we did in the course of the RCT evaluating the Graduation program in Ghana (Banerjee *et al.* 2020). The Graduation program itself increased labor supply; this is not particularly surprising because, while it made the household richer, it also gave household members a productive asset that opened up new opportunities for work. To dig deeper, we choose some of the households for an additional intervention to measure labor supply during hours not devoted to the (new) productive asset. These households were engaged in the production of cloth bags and paid a piece rate for every bag properly finished. This gave us a very credible way of measuring labor supply – we know exactly how many bags they produced and their exact quality (and less precisely, how long it took them to do the work). We also measured household earnings from working for others. Finally, we put some effort into measuring the inputs that they put into their own farms (including hired labor) and the resulting output.

When we compare those who were in the Graduation program with control households (both randomly chosen), we find that the former works the same number of minutes on bags as the latter but produce and therefore earn much more. In other words, despite having a productive asset to which they devote time, treated households also manage to put more effort into bag production (working more minutes than control on net). Since bag production requires no capital (we provide the capital) there is no reason that being in the graduation program would directly boost productivity. Therefore, the usual confound in interpreting the income effect, the liquidity effect described above, does not apply. Agricultural inputs also remain unaffected, meaning it is not the case that earnings from the Graduation program allow households to buy labor-substituting machinery for their farms and be more rested/productive.

In other words, the weight of the evidence clearly favors the view that being in the Graduation program reduces the perceived cost of working and perhaps the actual cost as well. This is reinforced by the fact that when households are given an unconditional cash transfer in a third intervention, they also do not work less. Indeed, their productivity in producing bags is higher, and their hours of work are no lower, though the differences are not statistically significant. One possible explanation for these findings is that the intervention makes households less anxious about their financial insecurities, as suggested by Mullainathan and Shafir (2013); another is that it makes them more forward-looking, as in Banerjee and Mullainathan (2010).

This example makes two separate points. First, that RCTs are in many ways ideal for building new theories because the experiment can be tailored to focus exactly on their key implications. In some cases, this is because the treatment can be exactly designed to pinpoint the key implication of the theory. A beautiful example is offered by Dean Karlan and Jon Zinman's "Observing Unobservables", (Karlan-Zinman 2009), which provides an experimental method for distinguishing between changes in selection and changes in the choice of action. In other cases, as in the case of the Ghana experiment, an experiment is powerful on account of being able to tailor measurement to the needs of theory.

The second point is that thinking through the implications of theories is extremely useful in setting up experiments and experimental measurement. Good experiments try to anticipate challenges to the interpretation of results. To do this well, it is very important to think through a model that permits a wide range of possibilities to avoid misinterpreting good news as bad news or vice versa. In a sense, while traditional empirical work tries to pare down the model for estimation, experimenters want relatively rich models to help articulate all that might falsify their interpretation of the experimental results.

4.2 Experiments to theory

The practice of working with the standard model (if for no other reason than to keep the estimation manageable) also limits the set of theories and hence the set of policies that we consider. The only available tools are ones that have a clear interpretation within the model. Some recent work of ours (Banerjee *et al.* 2017) illustrates why this may be a major limitation and highlights how the RCT approach circumvents this issue to test for possibilities outside of the standard model.

We start with the question: what is the best way to spread a piece of information that needs to quickly reach large numbers of people such that they can make the right choices for themselves? In particular, how might we best leverage social learning (the fact that people learn from each other and information spreads through a social network)? A lot of my early work studies the theory of social learning (Banerjee 1992, 1993) but I have more recently been involved in empirical studies of how the choice of seeds (the initial people who get informed) matters to how far (and fast) information spreads. This is of important policy concern, for example, when we wish to inform parents about an immunization camp to be held in the village in a few days, and to encourage them to have their kids immunized.

Our preliminary work on this topic was in the Indian state of Karnataka, where we asked villagers to name the best people for spreading a message about a fair or a music concert. Surprisingly, relatively few people got named (just 4 percent of the village), and those named were named on average by 9 separate sources. In other words, there is at least consensus on the best seeds. We called them "gossips". To verify that gossips are indeed more effective at spreading information, we ran an RCT in over two hundred villages. The ultimate goal was to spread information about a cell-phone raffle (still not something policymakers care about, but hold on). We varied, at the village level, whether the small number of initial seeds (3 or 5) were drawn from the set of gossips, the set of prominent people, or at random from the set of villages. Our results show that nearly 3 times as many people find out about the raffle if gossips are seeded.

We were then interested in whether this insight about gossips translated to a much more demanding setting, where the action we were trying to influence was immunization. One might worry that in such cases participation might be founded in much stronger preferences than in a cell-phone raffle. For this we worked with the government health department in some districts in the state of Haryana, where full immunization rates (i.e. the fraction of children who had got all five required shots) are among the lowest in the country. The goal was to get parents to bring their children to the immunization camps.

We once again followed the procedure for identifying gossips by asking villagers. To compare, especially given the more medical and perhaps controversial nature of immunizations, we also asked villagers to name the people they trusted and took a set of commonly trusted people as an alternative set of seeds. This was an instance where we slightly changed the experiment as part of trying to generalize the original insight – the idea was to set a more challenging benchmark based on the idea that trust plays an important role in who people listen to when it comes to important decisions like immunization. Iterating in this manner is an important part of how insights from RCTs get generalized.

We then ran a similar experiment across several hundred villages where the goal was now to increase the number of immunized children. We found that gossips convince twice as many additional parents to vaccinate their children as random seeds or "trusted" people. They are about as effective as giving parents a small incentive (in the form of cell-phone minutes) for each immunized child and thus end up costing the government much less.

Even though gossips proved incredibly successful at improving immunization rates, it is hard to imagine a policy of informing gossips emerging from conventional policy analysis. First, because the basic model of the decision to get one's children immunized focuses on the costs and benefits to the family (Becker 1981) and is typically not integrated with models of social learning. Indeed, work that empirically models the decision to pass on information within networks is itself at a very preliminary stage (see Banerjee *et al.* 2016). Perhaps as empirical network economics develops further there will be better integration between the two literatures.

The deeper problem is that the starting point of a conventional policy exercise is the best simple model that is currently available. The theory of social learning on networks does have a view on who would be best to inform (see for example Banerjee *et al.* 2013). The idea is to focus on those who are central to the network in an eigenvector sense, meaning those who are well-connected to well-connected people. It makes no mention of gossips or other motivated communicators. We therefore tried to infer from our experimental evidence how well we would have done if we had targeted the central people rather than the gossips. We found that while many gossips are central, on average, targeting gossips is more effective than targeting central players[7]. Given that knowing who is central requires mapping out the social network, which is something that is much more expensive and challenging than asking people to list gossips, it is very clear that using gossips is better.

It is not that one could not imagine a theory (and a related model that can be estimated) that emphasizes the role of motivated communicators. But given that the default understanding in conventional economics is that data is scarce, the tendency is to stick as closely as possible to the standard model, which discourages exploration. The attitude with RCTs is the opposite: the presumption is that it is not so hard to set up an experiment to test any single hypothesis (whether or not it is true) and therefore there is a lot of emphasis on coming up with interesting and original hypotheses (and by implication interesting and new models) to test. That is why getting to gossips is a natural outcome of doing RCTs.

Theories give shape to RCTs and RCTs often make us think of new theories, some of which would have never come out of conventional theorizing.

5. CONCLUSION

Finally why do we economists run RCTs? The answer should be obvious from the previous sections. I argue throughout the essay that extracting the right lessons from a set of RCTs is always of matter of strategically combining statistical methods with economic thinking, and that the nature of economic theorizing is transformed by the availability of results from RCTs. Economists are therefore both very well placed to design RCTs and to learn from them.

The useful question is therefore no longer "are RCTs economics?"

---

7. We go on to show theoretically that network members can learn who is central by observing the frequency with which someone's name gets mentioned in stories that come to them through the network.

Rather, it is "what is economics post-RCTs?" In other words, how does economics need to evolve to best take advantage of the vastly increased access to RCTs? Certainly, while there will probably always be more possible theories than credible facts, we need to adjust to the possibility of generating new facts when needed. Bad assumptions should not continue to be justified by the fact that we have no credible evidence against them. Perhaps a more radical thought is that we may want to abandon the ideal of a single, extremely spare, standard model that captures all relevant aspects of economic life. It may be more useful to build models with ingredients tailored to the particular context – specific types of behavioral assumptions that go beyond the "standard" model, specific assumptions about market failures, all based, as best as possible, on results from past research in similar settings. This is in effect what a lot of empirical researchers already do, but mostly on an ad-hoc basis, with the result that we continue to default to the standard model whenever our results do not directly conflict with it. To go back to the example of gossips, the fact that these few people are more keen on communicating than the rest of the network, is not yet standard convention while modeling networks. But perhaps it should be.[8] Finally, I think we can safely abandon the idea that RCTs are a minor diversion in the long arc of economics. All too many researchers (including many who were not brought up on RCTs) have sensed the possibilities that they offer. This Pandora's box cannot be closed.

REFERENCES

Andreoni, James, Michael A. Kuhn, and Larry Samuelson. 2019. "Building Rational Cooperation on Their Own: Learning to Start Small." *Journal of Public Economic Theory*, **21**(5): 812–825.

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

Attanasio, Orazio P., Costas Meghir, and Ana Santiago. 2012. "Education Choices in Mexico: Using a Structural Model and a Randomized Experiment to Evaluate PROGRESA." Review of Economic Studies, 79(1): 37–66.

Bandiera, Oriana, Robin Burgess, Narayan Das, Selim Gulesci, Imran Rasul, and Munshi Sulaiman. 2017. "Labor Markets and Poverty in Village Economies." *The Quarterly Journal of Economics,* **132**(2): 811–870.

Banerjee, Abhijit. 1992. "A Simple Model of Herd Behavior." *The Quarterly Journal of Economics,* **107**(3): 797–817.

8. Not that every model needs to capture every empirically relevant aspect of reality. Far from it. Good models have to be based on a judgment about the set of ingredients that are most likely to matter for a particular conclusion; what is important to have is the right universe of ingredients to from which to draw.

Banerjee, Abhijit. 1993. "The Economics of Rumors." *The Review of Economic Studies*, **60**(2): 309–327.
Banerjee, Abhijit. 2008. "Big Answers for Big Questions: The Presumption of Growth Policy." Prepared for the Brookings conference on "What Works in Development? Thinking Big and Thinking Small".
Banerjee, Abhijit, Rukmini Banerji, James Berry, Esther Duflo, Harini Kannan, Shobhini Mukerji, Marc Shotland, and Michael Walton. 2016. "Mainstreaming an Effective Intervention: Evidence from Randomized Evaluations of 'Teaching at the Right Level' in India." Working Paper.
Banerjee, Abhijit, Arun G. Chandrasekhar, Esther Duflo, and Matthew O. Jackson. 2013. "The Diffusion of Microfinance." *Science*, **341**: 363–370.
Banerjee, Abhijit, Arun G. Chandrasekhar, Esther Duflo, and Matthew O. Jackson. 2016. "Gossip: Identifying Central Individuals in a Social Network." NBER Working Paper 20422.
Banerjee, Abhijit, Sylvain Chassang, Sergio Montero, and Erik Snowberg. 2017. "A Theory of Experimenters." NBER Working Paper 23867.
Banerjee, Abhijit, and Esther Duflo. 2003. "Inequality and Growth: What Can the Data Say?" J*ournal of Economic Growth*, **8**(3): 267–299.
Banerjee, Abhijit, and Esther Duflo. 2004. "What Do Banks (Not) Do?" Massachusetts Institute of Technology Working Paper.
Banerjee, Abhijit, and Esther Duflo. 2011. *Poor Economics.* PublicAffairs. New York, NY.
Banerjee, Abhijit, and Esther Duflo. 2019. *Good Economics for Hard Times.* PublicAffairs. New York, NY.
Banerjee, Abhijit, Esther Duflo, Raghabendra Chattopadhyay, and Jeremy Shapiro. 2016. "The Long Term Impacts of a 'Graduation' Program: Evidence from West Bengal." Working Paper.
Banerjee, Abhijit, Esther Duflo, Nathanael Goldberg, Dean Karlan, Robert Osei, William Pariente, Jeremy Shapiro, Bram Thuysbaert, and Christopher Udry. 2015. "A Multifaceted Program Causing Lasting Progress for the Very Poor: Evidence from Six Countries." *Science*, **348**(6236): 772–788.
Banerjee, Abhijit, Esther Duflo, Daniel Keniston, and Nina Singh. 2019. "The Efficient Deployment of Police Resources: Theory and New Evidence from a Randomized Drunk Driving Crackdown in India." NBER Working Paper 26224.
Banerjee, Abhijit, Rema Hanna, Gabriel Kreindler, and Benjamin A. Olken. 2017. "Debunking the Stereotype of the Lazy Welfare Recipient: Evidence from Cash Transfer Programs." *The World Bank Research Observer,* **32**(2): 155–184.
Banerjee, Abhijit, and Sendhil Mullainathan. 2010. "The Shape of Temptation: Implications for the Economic Lives of the Poor." NBER Working Paper 15973.
Banerjee, Abhijit, and Andrew F. Newman. 1993. "Occupational Choice and the Process of Development." *Journal of Political Economy,* **101**(2): 274–298.
Becker, Gary. 1981. *A Treatise on the Family*. Harvard University Press. Cambridge, MA.
Benjamin, Dwayne. 1992. "Household Composition, Labor Markets, and Labor Demand: Testing for Separation in Agricultural Household Models." *Econometrica*, **60**(2): 287–322.
Dasgupta, Partha, and Debraj Ray. 1986. "Inequality as a Determinant of Malnutrition and Unemployment: Theory." *The Economic Journal,* **96**(384): 1011–1034.
Duflo, Esther. 2020. "Field Experiments and the Practice of Policy." Prize Lecture.
Duflo, Esther, Michael Greenstone, Rohini Pande, and Nicholas Ryan. 2013. "Truth-Telling by Third-Party Auditors and the Response of Polluting Firms:

Experimental Evidence from India." *The Quarterly Journal of Economics,* **128**(4): 1499–1545.

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

Gandelman, Néstor, and Rubén Hernández-Murillo. 2014. "Risk Aversion at the Country Level." Federal Reserve Bank of St. Louis Working Paper 2014-005B.

Karlan, Dean, and Jonathan Zinman. 2009. "Observing Unobservables: Identifying Information Assymetries With a Consumer Credit Field Experiment." *Econometrica*, **77**(6): 1993–2008.

Leibenstein, Harvey. 1957. *Economic Backwardness and Economic Growth*. Wiley. Hoboken, NJ.

Meager, Rachael. 2019. "Understanding the Average Impact of Microcredit Expansions: A Bayesian Hierarchical Analysis of Seven Randomized Experiments." *The American Economic Journal: Applied Economics,* **11**(1): 57–91.

Mullainathan, Sendhil, and Eldar Shafir. 2013. *Scarcity: Why Having Too Little Means So Much.* Henry Holt and Company. New York, NY.

Todd, Petra E., and Kenneth I. Wolpin. 2010. "Structural Estimation and Policy Evaluation in Developing Countries." *Annual Review of Economics*, **2**: 21–50.

Todd, Petra E., and Kenneth I. Wolpin. 2006. "Ex-Ante Evaluation of Social Programs." PIER Working Paper 06-022.

Wall Street Journal – Centennial Edition. June 23, 1989.