An Introduction to the "Handbook of Field Experiments"

Abhijit Vinayak Banerjee Esther Duflo

August 27, 2016

Many (though by no means all) of the questions that 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 (Fisher, 1925) proposed a method to answer such causal questions: Randomized Controlled Trials (RCTs) . In an RCT , the assignment of different units to different treatment groups is chosen randomly. This ensures that no unobservable characteristics of the units are 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 economics.

By many accounts, the first large-scale social experiment was the New Jersey Income Maintenance Experiment, which was initiated in 1968 and tested the impact of income transfers and tax rates on labor supply. The next few decades, as chapter 1 (Gueron, 2016) and chapter 18 (von Wachter and Rothstein, 2016) in this volume reminds us, were a sometime tortuous journey eventually leading to a more widespread acceptance of RCTs, both by policymakers and by academic researchers. While this acceptance first took hold in the US, starting in the mid 1990s it extended to developing countries, where the RCT "revolution" took the field by storm.

At this point, the method has gained widespread acceptance (though there continue to be vocal critics and active debates, many of which this Handbook covers), and there is now a large body of research on field experiments, both in developed and developing countries. We feel that we have collectively learnt an enormous amount from this literature, both in terms of how to conduct and analyze experiments, but also about the methodological contributions

experiments have made to economics and about the world. For this volume, we asked some of the foremost experts in the field to distill these learnings, as well as discuss the most important challenges and open questions for future work. In this short introduction, we provide what is our (admittedly personal, subjective, and somewhat biased towards our own field, development economics) assessment of the impacts that the past 20 years of field experiment research have had, both on how we do research and how we understand the world.

1 The impact on the way we do research

[1]

The remarkable growth in the number of RCTs is, in itself, a dramatic change in some fields. 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 are now involved in one or more randomized controlled trials (e.g. Daron Acemoglu, Derek Neal, Martin Ravallion, Mark Rosenzweig).

Early discussions of the merits (or lack thereof) of randomization put significant emphasis on its role in the reliable identification of internally valid causal effects and the external validity of such estimates. We and others have already had these discussions in various forum (Heckman, 1992; Banerjee et al., 2007; Duflo et al., 2007; Banerjee and Duflo, 2009; Deaton, 2010), and we will not reproduce them here. As we had also begun 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.

1.1 A greater focus on identification across the board

Starting with Neyman (1923) (who used it as a theoretical device) and Fisher (1925) (who was the first to propose physically randomizing units), the original motivation of randomized experiments was a focus on the credible identification of causal effects. As Imbens and Athey (2016) write in chapter 2 of this volume:

[1] This section draws on Banerjee, Duflo, and Kremer (2016)

> 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 view 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 largely on 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, most researchers in the social sciences 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, namely: matching, instrumental variables, difference-in-differences and regression discontinuity designs. Development economics also embraced these methods starting in the 1990s, but some researchers further decided that it may be possible to go straight to the "ideal" experiments (RCTs), and therefore researchers began to go back and forth between exper-

imental and non-experimental studies. This means that the experimental and non-experimental literatures developed in close relationship, constantly cross-fertilizing each other.

In development economics, 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 or at least well-defined 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, the standards of the non-experimental literature have therefore improved tremendously over the last few decades, without necessarily sacrificing its ability to ask broad and important questions. To highlight 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; Padró i Miquel, Qian, and Yao (2012) use a difference-in-differences strategy to study village democracy; and Banerjee and Iyer (2005) and Dell (2010) each 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, such as 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, clustering, non-compliance, attrition, spillovers and contamination, etc.). These methods are very clearly exposited in chapter 2 (Imbens and Athey, 2016) on the econometrics of experiments, and most chapters provide examples of their uses.

Structural methods are also increasingly combined with experiments to estimate counterfactual policies (See chapter 18 (von Wachter and Rothstein, 2016) for a number of examples from the developed world, as well as Todd and Wolpin (2006) and Attanasio, Meghir, and Santiago (2012) for developing country examples).

More recently, machine learning techniques have also been combined with experiments to model treatment effect heterogeneity (see chapter 2 (Imbens and Athey, 2016)).

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 that are not 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.

Interestingly, somewhat counter to this tendency to blur the boundaries between experiments and non-experiments, in chapter 2, Imbens and Athey (2016) provide a coherent framework for designing and analyzing experiments that puts randomization at the center:

> A major theme of the chapter is that we recommend using statistical methods that are directly justified by randomization, in contrast to the more traditional sampling-based approach that is commonly used in econometrics. In essence, the sampling based approach considers the treatment assignments to be fixed, while the outcomes are random. Inference is based on the idea that the subjects are a random sample from a much larger population. In contrast, the randomization-based approach takes the subject's potential outcomes (that is, the outcomes they would have had in each possible treatment regime) as fixed, and considers the assignment of subjects to treatments as random.

Thus, the methods they propose to analyze experiment sometimes differ from "traditional" econometrics: for example, instead of controlling for covariates (what researchers routinely do), which can easily lead to bias in finite sample, they suggest placing the data into strata, analyzing the within group experiments, and averaging the results. This is directly justified by the randomization of the treatment and does not require any additional assumptions. They also suggest doing as much as possible ex-ante through the design of the experiment to avoid any ex-post adjustment.

1.2 Assessing external validity

In the words of Imbens and Athey (2016) (chapter 2): "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. Other individuals could always be different and react differently to the treatment, and any future treatment could be ever so slightly different from what has been tested. As Banerjee, Chassang, and Snowberg (2016) (chapter 3) 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 and Duflo, 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 special, and non-representative of the future population that could be treated. Chapter 4 (Glennerster, 2016) provides a list of the characteristics of an ideal partner: they must have sufficient scale, be flexible and technically competent in the area of the program, have expertise and reputation, have low staff turnover, and possess a desire to know the truth. In other words, they are clearly not representative of the typical NGO or government, and this has clear implications on what can be generalized from those studies.

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 might 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 data that is presented to them with their own prior knowledge of their settings. From our experience, when presented with evidence from a program of interest, the immediate reaction of a policymaker is typically to ask whether an RCT

could be done in their own context.

There is one clear advantage that RCTs 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, also in principle, get a handle on how treatment effects might vary by context. Of course, if we allow the the world to vary in infinite ways, this is not sufficient to say anything much on its own. But there are several ways to make progress.

1.2.1 Combine existing evaluations and conduct meta-analyses

A first approach is to combine existing evaluations, and make assumptions about the possible distribution of treatment effects. There are a variety of ways to doing so, ranging from the explicitly parametric — 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 — to more non-parametric procedures, such as those based on revealed preference. Chapter 18 (von Wachter and Rothstein, 2016) contains an extensive discussion of the trade-offs between the various approaches in the context of the evaluation of social programs in developed countries. Chapter 12 (Fryer, 2016) provides a systematic meta-analysis of 196 RCTs in education in the US in three domains.

One issue that arises with trying to do any kind of meta-analysis is the access to an unselected sample of results from an unselected sample of studies. Since there is publication bias in economics, the worry is that the sample of published studies may not be representative of all the studies that exist; furthermore, since researchers have some flexibility in the analyses to run, the available results may themselves be selected. This is where another advantage of RCTs kicks in: since they have a defined beginning and end, they can in principle be registered. To this end, chapter 4 (Glennerster, 2016) discusses how the American Economic Association recently created a registry of randomized trials (www.socialscienceregistry.org), which listed over 800 entries as of August 10. The hope is that all projects are registered, preferably before they are launched, and

that results are clearly linked to their respective study, so that in the future meta-analysts can work from the full universe of studies. Chapter 4 (Glennerster, 2016) and chapter 3 (Banerjee, Chassang, and Snowberg, 2016) also have a useful exchange on the value to go further than registration and pre- analysis plan, where researchers lay out in advance the hypotheses to be tested and the regressions to be run.[2] Overall, both chapters point out the value in tying the hands of a partner who may be too eager to show success, but also emphasize that this comes with the cost of losing the flexibility to explore the data. In chapter 3, Banerjee, Chassang, and Snowberg (2016) point out that, if the data is available to others, there is in principle no reason to pre-specify a specific analysis, since anyone can decide what to run. This ties in to another issue that is discussed in chapter 4 (Glennerster, 2016): the need for open access of complete and usable data, both for reproducing existing analyses and for running others. This is an area where a lot of progress has been made, and hopefully more will be made in years to come.

1.2.2 Use other experiments to understand mechanisms

A second approach is to use the results from other experiments to test specific channels, and support the conclusions from the policy experiment. One way to do is to draw parallels between those results and results from laboratory experiments conducted in comparable settings (see chapter 9 (Gneezy and Imas, 2016)). Another option involves carrying out additional field experiments that provide support for the causal channels that underlie the policy claim (see chapter 14 (Kling, Ludwig, Congdon, and Mullainathan, 2016)).

1.2.3 Multi-site projects

A third 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, which is discussed in chapter 17 (Hanna and Karlan, 2016). BRAC, perhaps 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

[2]Paluck and Shafir (2016) also discuss the merit of pre-registration and pre-analysis plan for an experimenter who has some construal of what the results should be.

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 remained 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.

1.2.4 Structured speculation

One issue is that there is little the researcher can do ex-post to causally identify the source of differences in findings across countries. An option for multi-site projects 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 chapter 3 (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

According to Banerjee, Chassang, and Snowberg (2016), structured speculation has three advantages: 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 external validity. 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 idea is just being proposed, there are few examples as yet. A notable example is Dupas (2014), who 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.

Dupas (2014) then answers the question "For what types of health products and contexts would we expect the same results to obtain?" 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.

1.3 Testing theories

This discussion makes clear that the talking about external validity only makes sense once we understand the lesson that we want to generalize. Reflecting on the problem of partner selection that we mentioned earlier, in chapter 4, Glennerster (2016) writes:

> Whether we want to prioritize having a representative partner or a highly committed partner depends on the objective of the research. If we are testing an underlying

> human behavior—such as a willingness to pay now for benefits in the future—the representativeness of the partner may be less relevant. If we want to know whether a type of program, as it is usually implemented, works, we will want to prioritize working with a representative partner. Note that "does this type of program work" is not necessarily a more policy-relevant question than a more general question about human behavior. By their nature, more general questions generalize better and can be applied to a wider range of policy questions.

A big contribution of field experiments has been the ability to test theory. In chapter 2, Imbens and Athey (2016) argue "a randomized experiment is unique in the control that the researcher has over the assignment mechanism." 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 test implications of existing theories or 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 has mainly been to obtain causal effect of social policies that were often fairly comprehensive packages (Gueron, 2016), though according to chapter 14 (Kling, Ludwig, Congdon, and Mullainathan, 2016) there

has been a recent revival of what they call "mechanism experiments" which they define to be:

> ...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 are a specific version of experiments that test theories which distinctively have a relatively direct implication for the design of some policy.

Experiments that test theories, including mechanism experiments, have always had an important place in development economics and are now also used in developed countries. Banerjee and Duflo (2009) discuss some early examples of mechanism experiments including the justly influential papers on "observing unobservables" by Karlan and Zinman (2009). A number of these are discussed in chapter 7 (Bertrand and Duflo, 2016), chapter 11 (Dupas and Miguel, 2016), and chapter 17 (Hanna and Karlan, 2016).

Another area where it is now standard to use field experiments to test theories is in the growing literature on replicating tests of theories that were previously conducted in the laboratory in more realistic settings. Chapter 6 (Al-Ubaydli and List, 2016) and chapter 9 (Gneezy and Imas, 2016) are both excellent introduction to this literature, with the first focusing on theoretical predictions about market outcomes while the second is more about understanding preferences. By moving from the lab to the field, the studies that are reviewed in these two chapters aim to select a more relevant population, and to place them in situations that are not artificial, in order to test these theories in the contexts that are relevant in practice. The idea is that, in the lab, people do not behave as they would in reality. Chapter 5 (Paluck and Shafir, 2016) goes one step further in helping us think about how an experimenter must design an experiment to successfully

test a theory. They place the notion of "construal" at the center of their approach. They write: "Construal is defined as the individual's subjective interpretation of a stimulus, whether the stimulus is a choice set, a situation, another person group of people, or an experimental intervention." In order to successfully test a theory, the experiment must be designed such that the participants understand the world (and the different treatments) in the way the experimenter intended, and therefore their action and behavior in the different conditions can be interpreted. Of course, construal is relevant for other research as well (it affects how people will respond to a survey). But it is particularly important in an experimental set up, when a researcher is thinking about relevant manipulation. There is no magic recipe to do this, but Paluck and Shafir emphasize and encourage us to use this lens to think about basic experimental practice : piloting, as well as open-ended and open-minded observations, in the early phase of an experiment to ensure that the participants' construal is the same as the researchers; a "manipulation" check to make sure that participants understood that they were being treated; and decisions on whether to be present or not during an experiment.

1.4 Data collection

Data collection is at the core of experimental work, since administrative data is not always available or sufficient to obtain information on the relevant outcome. Considerable progress has been made on this front. Chapter 4 (Glennerster, 2016) gives specific and useful guidance on how researchers can insure the validity of the data that they have collected, summarizing best practices on monitoring, back checking, and effective use of information technology. Experiments have also spurred creativity in measurement, and Glennerster's chapter, as well as almost all the other chapters, covers these innovations. We elaborate a bit more on these issues here.

In principle, there is no automatic link between careful and innovative collection of microeconomic data and the experimental method. 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 then easier to afford than in the context

of a large multipurpose household or firm survey. By contrast, observational studies must often rely for identification on variation (policy changes, market-induced variation, natural variation, supply shocks, etc.) that cover large populations, requiring the use of a large dataset 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 ex post to do a sophisticated data collection exercise specifically targeted to the question, it is generally impossible to do it for the preprogram situation. This precludes the use of a difference-in-differences strategy for these types of outcomes, which again limits the incentives to collect them ex-post.

Some of the most exciting recent developments related to field experiments have to do with measurement. Researchers have turned to other sub-fields of economics as well as different fields altogether to borrow tools for measuring outcomes. Examples include soil testing and remote sensing in agriculture (see chapter 15 (de Janvry, Sadoulet, and Suri, 2016) for a review on agriculture); techniques developed by social psychologists for difficult to measure outcomes such as discrimination and prejudice – audit and correspondence studies, implicit association tests, Goldberg Experiments and List experiments (see chapter 7 (Bertrand and Duflo, 2016) for a review on 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 chapter 11 (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, and 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 under one specific condition, by encouraging greater participation in community

meetings. Rigol, Hussam, and Regianni (in progress) provide another example of innovative data collection practices. For their experiment, in order to accurately measure if and when people wash their hands, they designed soap dispensers that could track when the pump was being pushed and hired a Chinese company to manufacture them. 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. Administrative data are often at the core of the analysis of experiments in the US (see chapter 1 (Gueron, 2016) and chapter 18 (von Wachter and Rothstein, 2016)), and the more recent availability of tax data has allowed to examine long term impacts of interventions (Chetty et al., 2011, 2016). Recently, the practice has also spread to developing countries. 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; Olken, Khan, and Khwaja (2016) use administrative tax data from Pakistan; and Attanasio, Medina, and Meghir (2016) use unemployment insurance data to measure the long term effect of job training in Colombia.

Another increasingly important source of data comes from the use of lab-in-the-field experiments either as predictors of the treatment effect (e.g. commitment devices should help those who have self-control problems more than others) or as an outcome (e.g. cooperation in a public goods game as a measure of success in creating social capital). Chapter 9 (Gneezy and Imas, 2016) provides a number of examples, but also warns against blindly trusting lab-in-field experiments to unearth deep preferences—for example, behavior in a dictator game may not necessarily predict pro-social behaviors in real-life contexts.

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 goes beyond the traditional surveys, 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.

1.5 Iterate and build on previous research in the same settings

Another 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 raise more questions than they can actually answer. In particular, we are often left wondering why things turned out the way they did and how to change the intervention to make things (even) better.

This is where the ability to keep trying different interventions can be enormously valuable. Chapter 12 (Fryer, 2016) on education in the developed world is in part a history of such a quest. Fryer details the process of trying to figure out what actually works in closing the black-white achievement gap, describing the long line of experiments that failed to deliver or deliver enough and the slow accretion of learnings from successes and failures. Through this process, the main directions eventually became clear and he is able to conclude:

> These facts provide reason for optimism. Through the systematic implementation of randomized field experiments designed to increase human capital of school-aged children, we have substantially increased our knowledge of how to produce human capital and have assembled a canon of best practices.

We see a very similar process of dynamic discovery in chapter 8 (Gerber and Green, 2016) on the question of how to influence voter turnout. Marketing experiments also feature dynamically evolving treatments (see chapter 10 (Simester, 2016)), as do some agricultural experiments (see chapter 15 (de Janvry, Sadoulet, and Suri, 2016)).

1.6 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 conditional cash transfer (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 what are the elements that are key to its success. In recent years, a number of papers have looked "inside" CCTS, relaxing the conditionality and altering it in other ways which are discussed in chapter

17 (Hanna and Karlan, 2016). Hanna and Karlan also highlight the challenge of fully unpacking a program in the context of their discussion of the graduation program, mentioned above, which provides beneficiaries with the gift of an asset, as well as access to a savings opportunity, health services and information, life coaching and a small stipend. They write:

> 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.

As this paragraph implies, 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 so 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, 2007; Patel et al., 2010) Much remains to be learned not just in regards to the promise of such life-coaching components, but also 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. The evaluation of potential reforms of Indonesia's Raskin program by Banerjee et al. (2015), discussed in chapter 17 (Hanna and Karlan, 2016), is an example.

2 The impact on the way we think about the world

Whether or not the main point of a particular RCT was to test a theory, its results end up altering our theories about the world. While this is true of all credible empirical work, it is especially true of RCT results. This is because one advantage of RCTs and RCT-like natural experiments is that they do not rely on any theory for identification and therefore open the door to questioning even the most basic assumptions of the field. In this section, we list some of the areas where there are robust insights derived from the RCT literature, mostly building upon the material discussed in various chapters throughout this volume.

2.1 On the value of better human capital

The literature from health RCTs in developing countries (summarized in chapter 11 (Dupas and Miguel, 2016)) confirms what one would suspect: that serious ailments like HIV and malaria have large income/productivity consequences (this is based on the random assignment of scarce treatments). Dupas and Miguel also report on some RCTs that look at longer term outcomes for children who received health interventions in childhood. In some instances, such as deworming, there are striking long-term effects on, for example, earnings as an adult. The long-term follow-up of the Moving to Opportunity experiment in the United States, described in some detail in chapter 18 (von Wachter and Rothstein, 2016), has similarly large earnings positive consequences for those who benefitted from the move to a less poor neighborhood at young ages. Both chapters suggest that the magnitude of the long-term effects have not been explained fully, given the relatively small short-term effects.

Unfortunately, there seems to be very little else in the RCT literature on either health or education, either in the developed or the developing world, that can help us understand the

channels through which interventions at relatively young ages can have persistent and large effects. This remains an important area for future work.

2.2 On reforming education

The one very clear message from the RCT literature on education summarized in chapter 12 (Fryer, 2016) (for the developed world) and in chapter 13 (Muralidharan, 2016) (for the developing world) is that Teaching at the Right Level (TaRL) is perhaps the central ingredient of programs that succeed in helping the average school-age student perform substantially better. The idea behind the intervention is very simple: the student's specific deficiencies need to be identified and addressed even if they do not align with what he or she is expected to know at his or her age or grade. This might seem obvious but both chapters make the point that it is often precluded by the compulsions of school systems—in particular the need to keep up with the curriculum.

The right way to implement TaRL, however, differs across the two contexts. Fryer makes the case for expensive high intensity tutoring while Muralidharan describes the success of a number of low-cost interventions, where a limited amount of focused teaching by minimally trained volunteers seem to have had large positive effects. This difference could reflect, among other things, differences in the starting point (the kids in the developing world are so much further behind that it is easy to move them) or the fact that the right kind of low-intensity tutoring has not yet been arrived at in the developed countries.

It is also striking how many well-regarded interventions either do not work at all or give relatively weak positive results. These include various aspects of school infrastructure, student incentives, increasing the teacher-student ratio, standard teacher training/professional development, altering the teacher selection process, and perhaps most strikingly, school vouchers. Other interventions like computer-assisted learning seem to deliver mostly zero or negative results, but there are also a few large, significantly positive results, all from the developing world. The difference may come from the opportunity cost of the time—perhaps the alternatives to learning from the computer are worse in developing countries, were teachers are often completely disengaged and frequently missing. Another mixed bag is teacher incentives, where both Fryer and Muralidharan report a few very large positive effects and many small or zero effects. The reason for the variation may lie in the details of how the incentives were implemented or in the internal

culture or management of the school.

2.3 On the design of redistributive programs

Income and incentive (substitution) effects on labor supply are at the heart of the design of redistributive and social insurance programs. If these effects are strong and negative, the extent of possible income transfer may be severely limited and the constrained optimal insurance will tend to be very partial. Reassuringly, from the point of view of the efficiency of redistributive policies, the evidence from the developed world summarized in chapter 18 (von Wachter and Rothstein, 2016) suggests that both elasticities are negative but tend to be small (around 0.1). The evidence from the developing world, summarized in chapter 17 (Hanna and Karlan, 2016), finds in fact no clear evidence of negative income effects on labor supply. Interestingly, the unconditional income transfers seem to have no effect on labor supply, while the transfer of assets, such as in the so-called "graduation" programs, seems to encourage people to work harder, if anything.[3] It should be recognized however that these are impure income effects, since the assets potentially increased the marginal product of labor, though that still supports the case for redistribution. In addition, two recent review articles of the evidence from developing countries suggest that the additional income is often used to boost nutrition (Banerjee, 2016) and does not increase the consumption of temptation goods (Evans and Popova, 2014), which further reinforces the case for redistribution.

Given this, it is not surprising that the beneficiaries of a variety of asset transfer programs have been found to be durably better off as much as 5 years after they has ceased to have any contact with the program itself. It remains to be seen whether this is the effect of the asset transfer per se or the whole package which nudges beneficiaries to use their assets for long-run economic betterment.

On the flip side, the absence of strong incentive effects means that it is costly to use financial incentives to change behavior. Chapter 18 (von Wachter and Rothstein, 2016) summarizes the evidence on a range of social programs in developed countries which try to use incentives to alter the job search and job retention behavior of those at the margins of the labor market and find limited effects at best. The experience from conditional cash transfers (CCTs) in the less-developed world (as discussed in chapter 17 (Hanna and Karlan, 2016)) is a bit more varied;

[3] Banerjee et al. (2016) provide a summary of the evidence on income effects on labor supply.

most of the programs do alter behavior but, with some exceptions, the cost of doing so tends to be substantial.

2.4 On the design of incentives for public officials

A somewhat related literature that is mainly focused on developing countries (though there are echoes in chapter 12 (Fryer, 2016) and chapter 18 (von Wachter and Rothstein, 2016)) emphasizes the difficulty of using incentives to get better performance from public officials. This is the subject of a small but growing literature that is reviewed in chapter 16 (Olken, Pande, and Finan, 2016). The chapter starts by demonstrating that government employees are paid a premium in developing countries, which is not true in the developed world. Efficiency wage theory would suggest that this would make it easier to give incentives to government employees, but that does not seem to be the case. Incentives based on job termination are very rarely used and there is lots of prima facie evidence of delinquency by these well-paid officials, which has inspired a body of recent RCTs focused on trying to improve government performance by providing better incentives and other means. One main take-away from this literature is that it is difficult to design proper incentives for these officials (because of the risk of perverse responses) and perhaps even more difficult to make sure that these incentives are actually implemented.

2.5 On access to financial products

Given the success of asset transfer programs in raising earnings, the natural presumption is that improved access to reasonably priced credit would have similar effects. Yet, as chapter 17 (Hanna and Karlan, 2016) makes clear (see also chapter 15 (de Janvry, Sadoulet, and Suri, 2016)), there is essentially no support from RCTs for this view (this particular literature is almost exclusively focused on the developing world). Improving access to micro-credit, to take the obvious example, seems to have some effects on direction of consumer spending but no effects on earnings or even business earnings. This might be because the microcredit product is poorly designed, or because credit discourages risk-taking, or because the loan amounts are too small to permit the borrowers to invest in projects that earn high returns (or for a variety of other reasons), but the fact itself is striking.

On the other hand, in the case of agriculture, there is clear RCT evidence of positive impacts

on earnings from access to subsidized crop insurance (see chapter 15 (de Janvry, Sadoulet, and Suri, 2016)). The study by Karlan et al. (2014) on agriculture in Ghana (discussed in chapter 15) is especially striking in this context because it finds large investment and productivity effects from access to subsidized insurance, but no investment or productivity effects from a cash transfer. The authors interpret this as saying that these farmers are not credit constrained, however it is then not clear as to why these farmers do not invest and self-insure by borrowing and lending. It is true that self-insurance is not as good as getting insurance from the market, but the welfare loss seems small relative to the productivity gains. We believe that there exist important modeling issues here that have yet to be resolved.

2.6 On the demand for insurance and other prophylactic products

While insurance seems to be very useful to low income beneficiaries—who are happy to purchase insurance when it is highly subsidized, and change their behavior to take advantage of it—there is very little demand for it at the market price or anywhere close to it. This is true of both crop insurance (see chapter 15 (de Janvry, Sadoulet, and Suri, 2016)) and health insurance (see chapter 11 (Dupas and Miguel, 2016)). de Janvry, Sadoulet, and Suri suggest that this is in part because of a trust deficit between the insurer and the insured, who think that the insurer would refuse to pay ex post. However, Dupas and Miguel make the point that the same lack of demand is also seen in the case of most health protection goods—such as deworming pills, insecticide-treated bed-nets, and vaccination—suggesting that the problem may be more general.

One possibility is that there is not enough information about the efficacy of these products. While there is prima facie evidence of an information deficit, chapter 11 (Dupas and Miguel, 2016) finds the impact of providing information on the demand for healthcare to be quite mixed. The alternative is that the lack of demand is related to the widely-documented phenomenon of present bias: essentially the problem is that prophylactic products require the buyer to pay now and for uncertain future benefits. However, we are clearly some distance from a full resolution of the problem of demand and further research is clearly necessary.

2.7 On preferences and preference change

Deviations like these from the standard model of "rational" behavior in economic models are the inspiration of the three chapters: chapter 6 (Al-Ubaydli and List, 2016), chapter 9 (Gneezy and Imas, 2016), and chapter 10 (Simester, 2016), though from somewhat different angles. Al-Ubaydli and List explicitly take on the question of robustness of these deviations. In particular, they focus on whether or not these deviations survive strong incentives and long practice, both of which are characteristics of long-term market participants (of course this is not the only population of interest—mothers, for example, only need to vaccinate their children a few times in their lives). They conclude that while some of these deviations go away with practice or when properly incentivized, many of them are indeed quite robust—e.g. professionals are not necessarily less likely to deviate than students—and point out that despite this, many individual markets still deliver outcomes that are quite close to what the conventional equilibrium would predict.

Gneezy and Imas (chapter 9) have a somewhat different concern: Do we actually pick up the deep preference parameter we are looking for when we use the outcomes from lab-in-the-field experiments? For example, they conclude that:

> The results suggest that incentivized lottery experiments typically used to elicit risk attitudes lack predictive power over the unincentivized general survey questions in predicting relevant real-world behavior such as investment choices.

On the other hand, they find that the gender difference in competitiveness as measured by performance in games does correlate strikingly with how patriarchal the society is.

Chapter 10 (Simester, 2016) describes field experiments in marketing. The marketing field takes as given that consumers have biases and use simple heuristics to make decisions. A significant part of marketing effort goes into exploiting those to push the product. Moreover, there is advertising, which is in part directed towards altering preferences.

The experimental evidence described in chapter 10 (Simester, 2016) is in part about understanding the nature of people's heuristics and biases, how marketers respond to these heuristics and biases, and what kinds of advertising are most effective in changing preferences. There seems to be no general lesson other than the fact that many contextual seem to matter and therefore

experimentation is quite valuable. As a result, there are now dynamic models of targeted marketing where the specific intervention varies based on the past experience with that particular client or group of clients, and these models are tested using experimental methods. This is a very different approach from most of the field experiment literature, where the interventions to be tested are chosen based on priori thinking rather than experimentation. This is in part possible because in this age of high internet penetration and big data, marketing instruments (prices, offers, advertising, etc.) can be varied at a high frequency and the reaction to the changes can be tracked and processed immediately. This is obviously not always the case in other areas of economics, but it is worth thinking about how to design more experiments which follow the marketing model.

Finally, chapter 7 (Bertrand and Duflo, 2016) focuses on one specific kind of preferences: those that lead to prejudice and discrimination. The chapter starts by showing that there is robust experimental evidence of prejudice and discrimination, including self-discrimination based on audit studies, willingness to pay studies, and various psychological tools like IATs and Goldberg Paradox experiments. On the more difficult question of whether these are based on innate preferences rather than statistical discrimination, they find less clear-cut experimental evidence. However, the balance of the evidence taken together suggests that preferences do play an important role. The second part of the chapter describes the experimental evidence showing that these identity-based preferences (whether innate or induced) have significant negative consequences both for those who are viewed negatively as a result and for productivity in general. The final section then goes into the question of whether these preferences can be altered by an appropriate choice of interventions. This is perhaps where the experimental evidence is the most valuable and the scarcest. Laboratory work suggests that preference change is indeed possible, but too few convincing field studies have been conducted.

2.8 On the role of the community

Discrimination is of course one important reason why the structure of communities can have significant negative effects. However, there is now a large experimental literature that looks for positive effects. Chapter 15 (de Janvry, Sadoulet, and Suri, 2016) reports on the literature on learning from friends and neighbors in agriculture. Chapter 17 (Hanna and Karlan, 2016) discusses the possibility of using the community's knowledge about its members to identify the

poor. Chapter 11 (Dupas and Miguel, 2016), chapter 13 (Muralidharan, 2016), and chapter 16 (Olken, Pande, and Finan, 2016) all discuss the possibility of a specific type of collective action: using the community to monitor and incentivize local government officials. Our overall assessment of this evidence is that it is disappointing. There are few very successful examples, but in most cases there is surprisingly little transmission/use of collective knowledge or collective action. There are of course plausible explanations—many of which are mentioned in the chapters—but understanding why the community does not make use of the information and access it clearly remains an important agenda item for the future.

2.9 On getting people to vote

One form of collective action that many people do engage in is voting. In fact, so-called rational models of voting find it very difficult to explain why quite so many people vote. Given that, theory is unlikely to be a very good guide to the question of how to enfranchise even more people, especially from socially excluded groups. Starting from this observation, political scientists Alan Gerber and Don Green decided to take a radically empiricist approach to understanding how to influence turnout: they essentially organized a long series of RCTs where they tried out all the standard approaches and combinations thereof. This effort inspired a large and growing literature in political science which is detailed in chapter 8 (Gerber and Green, 2016). They summarize the main learnings from it in the following succinct paragraph:

> One is that encouragements to vote tend to be more effective when delivered in person than via direct mail or email. Another is that advocacy messages that give voters reasons to support or oppose a given candidate or cause tend not to increase turnout. Yet another is that messages that forcefully assert the social norm of civic participation are often highly effective at stimulating turnout, especially in low salience elections.

3 Conclusion

Overall, these chapters provide an incredibly rich overview of the remarkable progress that has occurred over the last 20 years in regards to field experimentation, reflecting both on advances and the issues that remain, as well as providing useful research tips and insights into what the

next steps should be. We hope that this Handbook provide guidance, identifies knowledge gaps, spurs further creativity and leads to research that continues to challenge our assumptions and help us understand the world better.

References

Al-Ubaydli, O. and J. List (2016). Field experiments in markets. In A. V. Banerjee and E. Duflo (Eds.), *Handbook of Field Experiments, forthcoming*, Chapter 6.

Alesina, A., P. Giuliano, and N. Nunn (2013). On the origins of gender roles: Women and the plough. *The Quarterly Journal of Economics 128*(2), 469–530.

Attanasio, O. P., C. Fernández, E. O. Fitzsimons, S. M. Grantham-McGregor, C. Meghir, and M. Rubio-Codina (2014). 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 349*, g5785.

Attanasio, O. P., A. Medina, and C. Meghir (2016). Long term impact of vouchers for vocational training: Experimental evidence for Colombia. *American Economic Journal Forthcoming*.

Attanasio, O. P., C. Meghir, and A. Santiago (2012). Education choices in Mexico: using a structural model and a randomized experiment to evaluate progresa. *The Review of Economic Studies 79*(1), 37–66.

Bandiera, O., R. Burgess, N. Das, S. Gulesci, I. Rasul, and M. Sulaiman (2013). Can basic entrepreneurship transform the economic lives of the poor? *IZA Discussion Paper 7386*.

Banerjee, A. V. (2016). Policies for a better-fed world. *Review of World Economics 152*(1), 3–17.

Banerjee, A. V., A. H. Amsden, R. H. Bates, J. N. Bhagwati, A. Deaton, and N. Stern (2007). *Making aid work*. MIT Press.

Banerjee, A. V., S. Chassang, and E. Snowberg (2016). Decision theoretic approaches to experiment design and external validity. In A. V. Banerjee and E. Duflo (Eds.), *Handbook of Field Experiments, forthcoming*, Chapter 3.

Banerjee, A. V., J. Das, and R. Hussam (2016). Improving the quality of private sector health care in West Bengal. *Forthcoming*.

Banerjee, A. V. and E. Duflo (2009). The experimental approach to development economics. *Annu. Rev. Econ 1*, 151–78.

Banerjee, A. V., E. Duflo, N. Goldberg, D. Karlan, R. Osei, W. Parienté, J. Shapiro, B. Thuysbaert, and C. Udry (2015). A multifaceted program causes lasting progress for the very poor: Evidence from six countries. *Science 348*(6236), 1260799.

Banerjee, A. V., E. Duflo, C. Imbert, S. Mathew, and R. Pande (2016). Can e-governance reduce capture of public programs? Experimental evidence from India's employment guarantee. *Mimeo*.

Banerjee, A. V., E. Duflo, and M. Kremer (2016). The influence of randomized controlled trials on development economics research and on development policy. *Mimeo MIT*.

Banerjee, A. V., R. Hanna, J. C. Kyle, B. A. Olken, and S. Sumarto (2015). The power of transparency: Information, identification cards and food subsidy programs in Indonesia. *National Bureau of Economic Research* (No. w20923).

Banerjee, A. V., R. Hanna, B. A. Olken, and G. Kreindler (2016). Debunking the stereotype of the lazy welfare recipient: Evidence from cash transfer programs worldwide. *Mimeo*.

Banerjee, A. V. and L. Iyer (2005). History, institutions, and economic performance: The legacy of colonial land tenure systems in India. *The American Economic Review 95*(4), 1190–1213.

Bertrand, M. and E. Duflo (2016). Field experiments in discrimination. In A. V. Banerjee and E. Duflo (Eds.), *Handbook of Field Experiments, forthcoming*, Chapter 7.

Bolton, P., J. Bass, T. Betancourt, L. Speelman, G. Onyango, K. F. Clougherty, R. Neugebauer, L. Murray, and H. Verdeli (2007). Interventions for depression symptoms among adolescent survivors of war and displacement in northern Uganda: a randomized controlled trial. *JAMA 298*(5), 519–27.

Bolton, P., J. Bass, R. Neugebauer, H. Verdeli, K. F. Clougherty, P. Wickramaratne, L. Speelman, L. Ndogoni, and M. Weissman (2003). Group interpersonal psychotherapy for depression in rural Uganda: a randomized controlled trial. *JAMA 289*(23), 3117–3124.

Chassang, S., G. Padró i Miquel, and E. Snowberg (2012). Selective trials: A principal-agent approach to randomized controlled experiments. *The American Economic Review 102*(4), 1279–1309.

Chetty, R., J. N. Friedman, N. Hilger, E. Saez, D. W. Schanzenbach, and D. Yagan (2011). How does your kindergarten classroom affect your earnings? Evidence from Project STAR. *The Quarterly Journal of Economics 126*(4), 1593–1660.

Chetty, R., N. Hendren, and L. F. Katz (2016). The effects of exposure to better neighborhoods on children: New evidence from the Moving to Opportunity experiment. *The American Economic Review 106*(4), 855–902.

de Janvry, A., E. Sadoulet, and T. Suri (2016). Field experiments in developing country agriculture. In A. V. Banerjee and E. Duflo (Eds.), *Handbook of Field Experiments, forthcoming*, Chapter 15.

Deaton, A. (2010). Instruments, randomization, and learning about development. *Journal of Economic Literature 48*(2), 424–455.

Dell, M. (2010). The persistent effects of Peru's mining mita. *Econometrica 78*(6), 1863–1903.

Duflo, E., R. Glennerster, and M. Kremer (2007). Using randomization in development economics research: A toolkit. *Handbook of Development Economics 4*, 3895–3962.

Dupas, P. (2014). Short-run subsidies and long-run adoption of new health products: Evidence from a field experiment. *Econometrica 82*(1), 197–228.

Dupas, P. and T. Miguel (2016). Impacts and determinants of health levels in low-income countries. In A. V. Banerjee and E. Duflo (Eds.), *Handbook of Field Experiments, forthcoming*, Chapter 11.

Dupas, P., J. Robinson, and R. Dizon-Ross (2016). Governance and the effectiveness of public health subsidies. *Forthcoming*.

Evans, D. K. and A. Popova (2014). Cash transfers and temptation goods: A review of global evidence. *World Bank Policy Research Working Paper* (6886).

Fisher, R. A. (1925). *Statistical Methods for Research workers*. Genesis Publishing Pvt Ltd.

Freedman, D. A. (2006). Statistical models for causation what inferential leverage do they provide? *Evaluation Review 30*(6), 691–713.

Fryer, R. (2016). The production of human capital in developed countries: Evidence from 196 randomized field experiments. In A. V. Banerjee and E. Duflo (Eds.), *Handbook of Field Experiments, forthcoming*, Chapter 12.

Gerber, A. and D. Green (2016). Field experiments on voter mobilization: An overview of a burgeoning literature. In A. V. Banerjee and E. Duflo (Eds.), *Handbook of Field Experiments, forthcoming*, Chapter 8.

Glennerster, R. (2016). The practicalities of running randomized evaluations: Partnerships, measurement, ethics, and transparency. In A. V. Banerjee and E. Duflo (Eds.), *Handbook of Field Experiments, forthcoming*, Chapter 4.

Gneezy, U. and A. Imas (2016). Lab in the field: Measuring preferences in the wild. In A. V. Banerjee and E. Duflo (Eds.), *Handbook of Field Experiments, forthcoming*, Chapter 9.

Gueron, J. (2016). The politics and practice of social experiments: Seeds of a revolution. In A. V. Banerjee and E. Duflo (Eds.), *Handbook of Field Experiments, forthcoming*, Chapter 1.

Hanna, R. and D. Karlan (2016). Designing social protection programs: Using theory and experimentation to understand how to help combat poverty. In A. V. Banerjee and E. Duflo (Eds.), *Handbook of Field Experiments, forthcoming*, Chapter 17.

Heckman, J. (1992). Randomization and social policy evaluation. In C. Manski and I. Garfinkel (Eds.), *Evaluating Welfare and Training Programs*. Cambridge: Harvard University Press.

Imbens, G. and S. Athey (2016). The econometrics of randomized experiments. In A. V. Banerjee and E. Duflo (Eds.), *Handbook of Field Experiments, forthcoming*, Chapter 2.

Karlan, D., R. Osei, I. Osei-Akoto, and C. Udry (2014). Agricultural decisions after relaxing credit and risk constraints. *Quarterly Journal of Economics 129*(2).

Karlan, D. and J. Zinman (2009). Observing unobservables: Identifying information asymmetries with a consumer credit field experiment. *Econometrica 77*(6), 1993–2008.

Kling, J., J. Ludwig, B. Congdon, and S. Mullainathan (2016). Social policy: Mechanism experiments and policy evaluations. In A. V. Banerjee and E. Duflo (Eds.), *Handbook of Field Experiments, forthcoming*, Chapter 14.

Kreindler, G. (In progress). Driving Delhi? The impact of driving restrictions on driver behavior. *Working Paper*.

Muralidharan, K. (2016). Field experiments in education in developing countries. In A. V. Banerjee and E. Duflo (Eds.), *Handbook of Field Experiments, forthcoming*, Chapter 13.

Neyman, J. (1923 [1990]). On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science 5*(4), 465–472.

Olken, B. A. (2007). Monitoring corruption: Evidence from a field experiment in Indonesia. *Journal of Political Economy 115*(2).

Olken, B. A., A. Q. Khan, and A. Khwaja (2016). Tax farming redux: Experimental evidence on performance pay for tax collectors. *Quarterly Journal of Economics 131*(1), 219–271.

Olken, B. A., R. Pande, and F. Finan (2016). The personnel economics of the state. In A. V. Banerjee and E. Duflo (Eds.), *Handbook of Field Experiments, forthcoming*, Chapter 16.

Padró i Miquel, G., N. Qian, and Y. Yao (2012). Social fragmentation, public goods and elections: Evidence from china. *NBER Working Paper No. 18633*.

Paluck, E. L. and E. Shafir (2016). The psychology of construal in the design of field experiments. In A. V. Banerjee and E. Duflo (Eds.), *Handbook of Field Experiments, forthcoming*, Chapter 5.

Patel, V., H. A. Weiss, N. Chowdhary, S. Naik, S. Pednekar, S. Chatterjee, M. J. De Silva, B. Bhat, R. Araya, M. King, et al. (2010). Effectiveness of an intervention led by lay health counsellors for depressive and anxiety disorders in primary care in Goa, India (MANAS): a cluster randomised controlled trial. *The Lancet 376*(9758).

Rao, G., F. Schilbach, and H. Schofield (In progress). Sleepless in Chennai: The economic effect of sleep deprivation among the poor. *Working Paper*.

Rigol, N., R. Hussam, and G. Regianni (In progress). Slipped my mind: Handwashing and habit formation. *Working Paper*.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology 66*(5), 688.

Rubin, D. B. (1981). Estimation in parallel randomized experiments. *Journal of Educational and Behavioral Statistics 6*(4), 377–401.

Simester, D. (2016). Field experiments in marketing. In A. V. Banerjee and E. Duflo (Eds.), *Handbook of Field Experiments, forthcoming*, Chapter 10.

Todd, P. E. and K. I. Wolpin (2006). Assessing the impact of a school subsidy program in Mexico: Using a social experiment to validate a dynamic behavioral model of child schooling and fertility. *The American Economic Review 96*(5), 1384–1417.

von Wachter, T. and J. Rothstein (2016). Social experiments in the labor market. In A. V. Banerjee and E. Duflo (Eds.), *Handbook of Field Experiments, forthcoming*, Chapter 18.