# The Experimental Approach to Development Economics

Abhijit V. Banerjee and Esther Duflo[1]

Massachusetts Institute of technology, Department of Economics and Abdul Latif Jameel Poverty Action Lab

Abstract:

Randomized experiments have become a popular tool in development economics research, and have been the subject of a number of criticisms. This paper reviews the recent literature, and discusses the strengths and limitations of this approach in theory and in practice. We argue that the main virtue of randomized experiments is that, due to the close collaboration between researchers and implementers, they allow the estimation of parameters that it would not otherwise be possible to evaluate. We discuss the concerns that have been raised regarding experiments, and generally conclude that while they are real, they are often not specific to experiments. We conclude by discussing the relationship between theory and experiments.

The last few years have seen a veritable explosion of randomized experiments in development economics and with it, perhaps inevitably, a rising tide of criticism. Almost all of the criticism is well-meant, recognizing the benefits of such experiments while suggesting that we not forget that there are a lot of important questions that randomized experiments cannot answer. Much of it is also not new. Indeed, most of the standard objections (and some not so standard ones) may be found in a single seminal piece by James Heckman, written over a decade ago (Heckman, 1992).

Much of this criticism has been useful, even when we do not entirely agree with it, both in helping us think through the strengths and limitations of what has been done, and in clarifying where the field needs to go next. However, we will argue that much of this criticism misses (or at least insufficiently emphasizes) the main reasons why there has been so much excitement surrounding experimental research in development economics. We will then return to the various criticisms, in part to clarify and qualify them, and in part to argue that, because of an imperfect recognition of what is exciting about the experimental agenda, there is a tendency to set up false oppositions between experimental work and other forms of research.

[1] We thank Guido Imbens for many helpful conversations.

## 1. The promise of experiments

Experimental research in development economics, like earlier research in labor economics and health economics, started from a concern about the reliable identification of program effects in the face of complex and multiple channels of causality. Experiments make it possible to vary one factor at a time and therefore provide "internally" valid estimates of the causal effect. The experimental work in the mid 1990s (e.g. Glewwe, Kremer and Moulin, forthcoming; Glewwe, Kremer, Moulin and Zitzewitz,2004; Banerjee, Jacob and Kremer, 2005), was aimed at answering very basic questions about the educational production function: does better access to inputs (textbooks, flipcharts in classes, lower student-teacher ratios) matter for school outcomes (attendance, test scores) and if so, by how much?

This research produced a number of surprising results. Improving access to textbooks from one per four or more students to one per every two does not affect the average test score (Glewwe, Kremer and Moulin, forthcoming); nor does halving the teacher-student ratio (Banerjee, Jacob and Kremer, 2005). On the other hand, one might also get surprisingly positive results: A study of treatment for intestinal worms in schools in Kenya (Miguel and Kremer, 2004), showed that a deworming treatment that costs 49 cents per child per year can reduce absenteeism by one-quarter. In part, this is because of externalities: worms are transmitted by walking barefoot in places where other children who are infected by worms have defecated. As a result, in terms of increasing attendance, deworming is nearly twenty times as effective as hiring an extra teacher (the cost for an extra child-year of education was $3.25 with deworming, as against around $60 for the extra teacher program, despite the fact the extra teacher was paid only $25 or so a month), even though both "work" in the sense of generating statistically significant improvements.

What this research was making clear is that at the level of the efficacy of individual ingredients of the educational production function, our intuition (or economic theory per se) was unlikely to help us very much—how could we possibly know, a priori, that deworming is so much more effective than hiring a teacher. More generally, a bulletin of the Abdul Latif Jameel Poverty Action lab compares the cost per extra child-year of education induced across an array of different strategies (J-PAL, 2005). The costs vary widely, between $3.50 per extra child-year for

deworming to $6,000 per extra child-year for the primary education component of PROGRESA, the Mexican Conditional Cash Transfer Program. Some of these programs (such as the PROGRESA programs) may have other objectives as well. But for those whose main goal is to increase education, it is clear that some are much cheaper than others. Even excluding PROGRESA, the cost per extra year of education induced range from $3.25 to over $200. Thus, even when comparing across programs to achieve the same goal, the rates of returns of public investment are far from being equalized.

Moreover, it became clear that economists were not the only people who were clueless--the implementing organizations were not much better informed. For example, the NGO that financed the deworming intervention was also initially enthusiastic about giving children school uniforms, though a randomized evaluation showed that the cost of an extra child year coming giving children a free uniform worked out to be $100 per child year.

Several important conclusions emerged from this experience. First, effective policy-making requires making judgments about the efficacy of individual components of programs, without much guidance from a priori knowledge. Second, however, it is also difficult to learn about these individual components from observational (i.e. non-experimental) data. The reason is that observational data on the educational production function often comes from school systems that have adopted a given “model”, which consists of more than one input. The variation in school inputs we observe therefore comes from attempts to change the model, which, for very good reasons, involves making multiple changes at the same time-. A good example is “Operation Blackboard” in India (Chin, 2005), a program of school upgrading which involve simultaneously hiring new teachers and providing teaching-learning material to the schools. Subsequent education programs in India (the District Primary Education Program, the Sarva Siksha Avian) have all had this feature. While there are important exceptions (for example, the fact that class size changes discontinuously with enrollment in Israel allows for a clean evaluation of the impact of just class sizes, see Angrist and Lavy (1992)), this means that a lot of the policy relevant knowledge that requires observing the effects of variation in individual components of a package may not be available in observational data. This is a first motivation for experiments.

One of the immediate implications of this observation is that, given the fixed cost of organizing an experiment and the fact that experiments necessarily require some time when program implementation has to be slowed down (in order to make use of the results), it is worth doing multiple experiments at the same time on the same population, which evaluate alternative

potential variants of the program.  For example, the World Bank provided money to school committees to hire extra teachers on short contracts to reduce class size in grade 1 in Kenya. When they worked with the school system to set up an evaluation of the program, the researches did not just assign the entire program to the randomly selected treatment schools (Duflo, Dupas and Kremer, 2008). Two additional dimensions of variation were introduced: training of the school committee that received the money to monitor the extra teacher, and tracking by prior achievement. This design thus allows estimating the impact of class size reduction without change in pedagogy, the relative merit of young, extra teachers on short contracts versus of regular, experienced, civil servant teachers, the role that suitably empowered school committees can play, and the impact of tracking by achievement in primary school. As in Banerjee, Jacob and Kremer (2005), albeit in a very different context, the study does not find that reducing class size without any other changes has a significant impact. However, it showed a strong positive impact of switching from the regular teacher to a contract teacher, a positive and significant impact of class size reduction when coupled with school committee empowerment and, for a given class size, strong benefit of tracking students, both for the weaker and the stronger students.

Other “multiple treatment experiments” include Banerjee, Cole, Duflo and Linden (2007), (remedial education and computer assisted learning), Duflo, Dupas, Kremer and Sinei (2006) and Dupas (2007), (various HIV-AIDS prevention strategies among adolescents), Banerjee, Barnerji, Duflo, Glennerster and Khemani (2008) (information and mobilization experiments in primary schools in India), Banerjee, Duflo, Glennerster and Kothari (2008) (demand and supply factor in improving immunization rates in India), Gine, Karlan and Zinman (2008) (two strategies to help smokers quit smoking), and many others.

A related observation is that from the point of view of building a useable knowledge base, there is a need for a process of dynamic learning: First because experimental results are often surprising and therefore require further clarification. Duflo, Kremer and Robinson (2008a,b) reflects exactly such an iterative process, where a succession of experiments on fertilizer use were run over a period of several years, each results prompting the need to try out a series of new variation in order to better understand the results of the previous one.

Second, from the point of view of optimal learning, it is often worth testing a broad intervention first to see whether there is an overall effect and then, if it is found to work, delving into its

individual components, as a way to understand what part of the broad program works.[2] Policy experiments often stops at the first step: one example is the popular PROGRESA-Opportunidades program in Mexico, which combined a cash transfer to poor families conditional on "good behavior" (investments in education and preventive health), with transfers to women, and some upgrading of education and health facilities. The program has been replicated in many countries, often along with a randomized evaluation. But it is only in an ongoing study in Morocco that different treatment groups are formed and compared, in order to evaluate the importance of the much-praised conditionalities. In this experiment, one group of villages receives a purely unconditional transfer, one group receives a "weak conditionality" transfer , where attendance requirements are only verified by teachers, and two groups receive a stricter variants of the conditionality (in one group, children attendance is supervised by inspectors; in the other, it is verified daily with a fingerprint recognition device).

While all this seems obvious in retrospect, it was only after the experience of the first few experiments that both researchers and the implementing organizations that they worked with fully appreciated what it all meant for them. From the point of view of the organizations it became clear that there was value in setting up relatively long-term relationships with researchers, so that the experimentation could constitute a process of on-going learning and multiple experiments of mutual interests could be designed. In other words, there was less emphasis on one-off evaluations, where the researcher is brought in to evaluate a specific program that the organization has already decided to evaluate. This is a difference with the evaluation literature in the US or Canada where, with a few important exceptions (e.g. Angrist, Lang and Oreopoulos, forthcoming) the programs to be evaluated are mainly chosen by the implementing agencies, and the researchers are evaluators.

From the point of view of the researchers, this offered the possibility of moving from the role of the evaluator to the role of a co-experimenter, with an important role in defining what gets evaluated. In other words, the researcher was now being offered the option of defining the question to be answered, drawing upon his knowledge of what else was known and the received theory. For example, Seva Mandir, a NGO in Rajasthan, India with whom Banerjee and Duflo had had a long standing relationship, was interested improving the quality of their

[2] Or the opposite: going from one intervention at a time to the full package makes sense when your priors are that some combination will work, while the opposite is better when you are generally skeptical.

informal schools. Their initial idea was to implement a teacher incentive programs based on test scores. However, they were persuaded by the results from Glewwe, Ilias and Kremer (2003) that a danger with teacher incentives would be teaching to the test or other short run manipulation of test scores. They then decided to implement an incentive program based on teacher presence. To measure attendance, in very sparsely populated area where schools are difficult to access, Duflo and Hanna proposed to use cameras with date and time stamps. While Seva Mandir was initially somewhat surprised by the suggestion, they agreed to try it out. In program schools (the "camera schools"), teachers took a picture of themselves and their students twice a day (morning and afternoon), and their salary was computed as a (non-linear) function of the number of days they attended. The results, reported in Duflo, Hanna and Ryan (2007) were quite striking: teacher absence dropped by from 40 percentage points to 20 percentage point, and students' performance also improved. Seva Mandir was convinced by these results, and decided to continue the program. However, they did not give up on the hope of improving the teachers' intrinsic motivation. Instead of extending the camera program in all their schools immediately, they thus decided to continue it in the schools where they had already been introduced, and spend some time experimenting with other programs, both in schools with cameras and in schools without. With Sendhil Mullainathan, They brainstormed about ways to motivate teachers. One idea was to give every child a diary to write in every day based on work done in school. On days where the student or the teacher was absent, the diary was to remain blank or be crossed out. Parents were supposed to look at the diary every week. The hope was that they would register just how much teacher and child absence was there. This, it turned out, did not succeed: parents apparently started from such a low opinion of school that the diary tended to persuade them that something was happening---parents have a higher opinion of diary schools than non-diary schools, and there was no impact on teacher presence. However, the diaries were very popular both with students and teachers, and induced teachers to work harder when present. Test scores improved in the diary schools. It thus appears that the diaries failed as a tool to improve teacher presence, but succeeded as a pedagogical tool. However, since this was not a hypothesis put forward in the initial experimental design, it may just be a statistical accident. Thus, while Seva Mandir will now put cameras in all schools (after several years, they continue to have a large impact on presence and tests scores), they will conduct a new diary experiment to see if the results on pedagogy persist.

One important consequence of this process has been the growing realization in the research community that the most important element of the experimental approach may lie in the power,

when working with a friendly implementing partner, to vary individual elements of the treatment in a way that helps us answer conceptual questions (albeit policy relevant ones) that could never be reliably answered in any other way.[3] One telling example is Berry (2008). While incentives based on school participation and performance have become very popular, it is not clear whether the incentives should target children (as in the programs evaluated in Angrist, Lang and Oreopoulos (2008) and Angrist and Lavy (2002)) or parents (as in Kremer, Miguel and Thornton (2007)). If the family were fully efficient the choice of the target should not make a difference, but otherwise it might. To answer this question Berry worked with Pratham in the slums of Delhi to design a program where students (or their parents) were provided incentives (in the form of toys or money) based on child's improvement in reading. He found that for initially weak students, rewarding the child is more effective in terms of improving test scores than rewarding the parents. Clearly without being able to vary who receives the incentives within the same context and in the same experiment, this study would not have been possible.

Experiments are thus emerging as a powerful tool for testing theories in the hands of those with sufficient creativity. Karlan and Zinman (2005) is one example. The project was conducted in collaboration with a South African lender that gives small loans to high risk borrowers at high interest rates. The experiment was designed to test the relative weights of ex post repayment burden (including moral hazard) and ex ante adverse selection in loan default. Potential borrowers with the same observable risk are randomly offered a high or a low interest rate in an initial letter. Individuals then decide whether to borrow at the solicitation's "offer" rate. Of those that apply at higher rate, half are randomly offered a new lower "contract" interest rate when they actually given the loan, while the remaining half continue at the offer rate. Individuals did not know ex ante that the contract rate could differ from the offer rate. The researchers then compared repayment performance of the loans in all three groups. The comparison of those who responded to the high offer interest rate with those who responded to the low offer interest rate in the population that received the same low contract rate allows the identification of the adverse selection effect, while comparing those who faced the same offer rate but differing contract rates identifies the repayment burden effect.

---

[3] While the constraint of working with an implementation organization does limit the set of questions you can ask, relative to what one can do in a lab experiment, the extra realism of the setting seems to be an enormous advantage.

The study found that women exhibit adverse selection but men exhibit moral hazard. The fact that this difference was unexpected poses a something of a problem for the paper (is it s statistical fluke, or a real phenomenon) but its methodological contribution is undisputed. The basic idea of varying prices ex post and ex ante to identify different parameters has since then been replicated in several different studies. Ashraf, Berry and Shapiro (2007) and Cohen and Dupas (2007) exploit it to understand the relationship between the price paid for a health protection good and its utilization. Raising the price could affect usage through a screening effect (those who buy at a higher price care more) or a "psychological sunk cost effect". To separate these effects, they randomize the offer price as well as the actual paid price. The effect of the offer price keeping the actual price fixed identifies the screening effect while the variation in the actual price (with a fixed offer price) pins down the suck cost effect. Ashraf et al (2007) study this for a water-purification product while Cohen and Dupas (2007) focus on bednets. In neither study is there much evidence of a psychological sunk cost effect. The experimental variation was key here, and not only to avoid bias: in the world we are unlikely to observe a large number of people who face different offer prices but the same actual price. These types of experiments are reminiscent of the motivation of the early social experiments (such as the negative income tax experiments), which aimed to obtain distinct wage and income variations to estimate income and substitution effects, which were not available in observational data (Heckman, 1992).

Other examples of this type of work are the experiments designed to asses whether there is a demand for commitment products: these products could be demanded by self-aware people with self-control problems. Ashraf, Karlan and Yin (2006) worked with a microfinance institutions in the Philippines to offer to their clients a savings product that let them chose to commit not to withdraw the money before a specific time or amount goal was reached. Gine, Karlan and Zinman (2008) worked with the same organization to invite smokers who want to quit to put a "contract" on themselves: money in a special savings account would be forfeited if they fail a urine smoking tests after several weeks. In both cases these were designed by the economists to solve a real world problem, but came with a strong theoretical motivation. The fact that these were new ideas that came from researchers made it natural to set up a randomized evaluation: since they were experimental in nature, the partners were typically happy to first try them out with a subset of their clients/beneficiaries.

These two sets of examples are focused on individual behavior. Experiments can also be set up to understand the way institutions function. An example is Bertrand, Djankov, Hanna and Mullainathan (forthcoming), who set up an experiment to understand the structure of corruption in process of obtaining a driving license in Delhi. They recruit people who are aiming to get a driving license, and set up three groups, one which receives a bonus for obtaining a driving license fast, one that gets free driving lessons, and a control group. They find that those in the "bonus" group do get their licenses faster, but those who get the free driving lessons do not. They also find that those in the bonus group are more likely to pay an agent to get the license (who, they conjecture, in turn bribes someone). They also find that hiring an agent is correlated with a lower probability to have taken a driving test before getting a driving license and to be able to drive. While they do not appear to find that those in the bonus group who get licenses are systematically less likely to know how to drive than those in the control group (which would be the litmus test that corruption does result in an inefficient allocation of driving licenses), this experiment provides suggestive evidence that corruption in this case does more than "grease the wheels" of the system.

The realization that experiments are a readily available option has also spurred creativity in measurement. While there are plenty of experiments which make use of standard methods, and plenty of non-experimental papers which have invested a lot in measurement (e.g. Olken (2007b) on measuring bribes, Manski (2004, and many other papers), on measuring expectations, the biological samples in the Indonesian Family Life Survey and the Health and Retirement Surveys, etc. ), the advantage that experiments offer 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 in order to assess the impact of the program does not have to be very large, and that data is typically collected especially for the purpose of the experiment. Elaborate and expensive measurement of outcomes is then possible.

By contrast, most quasi experimental observational studies rely on some kind of a change in policy for identification. These policy changes usually cover large populations, requiring the use of large data sets, often not collected for this specific purpose. Moreover, even if it is possible ex post to do a sophisticated data collection exercise specifically targeted to the program, it is generally impossible to do it for the pre-program situation. This precludes the use of a difference-in-differences strategy for these types of outcomes.

One example of the kind of data that was collected an experimental setting is Olken (2007a). 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 actually used. He then compared that with the level of material reported to be used. The difference is a measure of how much of the material was stolen, or never purchased but invoiced, and thus an objective measure of corruption. Olken then demonstrated that this measure of "missing inputs" is affected by the threat of audits, but not, except in some circumstances, by encouraging greater attendance at community meetings.

Another example of innovative data collection is found in Beaman, Chattophadhyay, Duflo, Pande and Topalova (2008). The paper evaluates the impact of mandated political representation of women in village councils on citizens' attitude towards women leaders. This is a natural randomized experiment in the sense that villages were randomly selected (by law) to be "reserved for women": in the "reserved" villages, only women could be elected as village head. To get a measure of "taste" for women leaders that would not be tainted by the desire of the respondent to please the interviewer, the paper implements "implicit association tests", developed by psychologists (Banaji, 2001). While those tests are frequently used by psychologists, and their use has also been advocated by economists (Bertrand, Chugh and Mullainathan, 2005) they had not been implemented in a field setting in developing country, and there had been almost no studies investigating whether these attitudes are "hard wired" or can be affected by features of the environment. The study also used another measure of implicit bias towards women, inspired by political scientists The respondents listen to a speech, supposedly given by a village leader, delivered either by male or female voice, and are asked to give their opinion of it. Respondents are randomly selected to receive either the male or the female speech. The difference in the ratings given by those who receive male versus female speeches is a measure of statistical discrimination. The paper then compares this measure of discrimination across reserved and un-reserved villages.

These are only two examples of a rich and creative literature. Many field experiments embed small lab experiments (dictator games, choices over lotteries, discount rate experiments, public good games etc.). There are innovations there too: for example, in ongoing research, in order to measure "social capital", Erica Field and Rohini Pande distributed lottery tickets to respondents, and gave the subjects the option to share them with members of their groups.

## 2. The concerns about experiments

As we mentioned, the concerns about experiments are not new. However many of these are based on comparing experimental methods, implicitly or explicitly, with other methods for trying to learn about the same thing. The message of the previous section is that the biggest advantage of experiments may be that they take us into terrain where observational approaches are not available. In such cases, the objections raised by critics of the experimental literature are best viewed as warnings against over-interpreting experimental results. There are however also cases where both experimental and observational approaches are available in relatively comparable forms, where there is, in addition, the issue of which approach to take. Moreover there are concerns about what experiments are doing to development economics as a field. The rest of this section lists these objections and then discusses them one by one.

### 2.1 Environmental dependence

Environmental dependence is a core element of generalizability. It asks the question, would we get the same result if we carry out the same experiment in a different setting or, more exactly, would the program that is being evaluated have the same effect if it was implemented elsewhere (not in the context of an experiment).

This is actually two separate concerns: First and most obviously, we may worry about the impact of differences in the experimental environment on the effectiveness of the program. One virtue of experiments is that they allow us to evaluate the mean effect of the program for a specific population without assuming that the effect of the program is constant across individuals. But if the effect is not constant across individuals, it is likely to vary systematically with covariates. For example, school uniforms will surely not have the same impact in Norway (where every child who needs one, no doubt, has one) that it has in Kenya. The question is where to draw the line: Is Mexico more like Norway or more like Kenya? The same issue also arises within a country. Clearly a priori knowledge can only help us here to some extent---simple economics suggests that uniforms will only have an effect in populations where the average wage is not too high relative to the price of uniforms, but how high is too high? If our theories are good enough to know this, or we are willing to assume that they are, then we probably do not need experiments anymore: theory may then be good enough to give us a sense of who tends to get a uniform, and who does not, and we could use this restriction to convincingly

estimate structural models of the impact of school uniforms. In other words, without assumptions, results from experiments cannot be generalized beyond their context; but with enough assumptions, observational data may be sufficient. To argue for experiments, we need to be somewhere in the middle.

A second issue comes from worrying about implementer effects. In particular, the smaller the implementing organization, the greater the concern that the estimated treatment effect reflects the unique characteristics of the implementer. A related concern expressed by Heckman (1992) is that sites or organizations that accept to be part of an experiment may be different from others. For example, he points out that several sites refused to participate in the JTPA experiments, because they objected to randomization.

This problem can be partially mitigated by providing detailed information about the implementation in the description of the evaluation, emphasizing the place of the evaluated program within the overall action plan of the organization (how big was the evaluated piece relative to what they do, how was the implementing team selected, what decided the choice of location, etc.). Clearly for the results to be anything more than an initial “proof of concept”, the program must come from a program that is sufficiently well-defined and well-understood that its implementation routinely gets delegated to a large number of more or less self-sufficient individual implementing teams.

All this is however very loose and highly subjective (what is large enough? how self-sufficient?, etc.). To address both concerns about generalization, actual replication studies need to be carried out. Additional experiments have to be conducted in different locations, with different teams. If we have a theory that tells us where the effects are likely to be different, we focus the extra experiments there. If not, we should ideally choose random locations within the relevant domain.

Indeed there are now a number of replication studies, although as Heckman pointed out, locations where experiments are run are generally not chosen randomly. The supplemental teaching (“balsakhi”) program evaluated by Banerjee et al. (2007), was actually deliberately carried out simultaneously in two separate locations (Mumbai and Vadodara) working with two separate implementing teams (both from the Pratham network, but under entirely separate management). The results turned out to be broadly consistent. Similarly Bobonis, Miguel and Sharma (2006) get similar impact of a combination of deworming and iron supplementation on school attendance in North India that Miguel and Kremer (2004) found in Kenya, and Bleakley

(2007) finds similar results using natural data from the US south in the early part of the 20th century using a natural experiment approach. The PROGRESA/Opportunidades program was replicated under different names and with slight variations in many countries, and in several of them, it was accompanied by a randomized evaluation (Colombia, Nicaragua, Ecuador, and Honduras; Morocco is under way). The results were very consistent across countries.

Other results turn out not to be replicable: An information campaign that mobilized parent's committees on issues around education and encouraged them to make use of a public program that allows school committees to hire local teachers where the schools are over-crowded, had a positive impact on learning outcomes in Kenya but not India (Banerjee, Banerji, Duflo, Glennerster and Khamani et al, 2008; Duflo, Dupas Kremer, 2008). And a similar intervention that sought to energize Health Unit Management Committees in Uganda reported a massive impact on hard to affect outcomes like infant mortality (Bjorkman and Svensson, 2007).

In addition to pure replication, cumulative knowledge is generated from related experiments in different contexts. Kremer and Holla's (2008) analytical review of 16 randomized experiment of price elasticity in health and education is a nice example. We will come back to these results in more detail below, but the key point here is that these experiments cover a wide range education and health goods and services, in several countries. A very strong common thread is the extremely high elasticity of the demands for these goods relative to their price, especially around zero (both in the positive and negative direction). While they are not strictly replications of each other, this clearly shows the value of cumulative knowledge in learning about one phenomenon.

It is clear, however, that there needs to be much more such replication research. Some worry that there are little incentives in the system to carry out replication studies (since journals may not be as willing to publish the fifth experiment on a given topic as the first one), and funding agencies may not be willing to fund them either. The extensive use of experiments in economics is still recent, so we do not know how big problem this might be, though given the many published estimates of the returns to education, for example, we are not too pessimistic. The good news is that several systematic replication efforts are underway. For example a program of asset transfers and training targeted to the ultra poor, originally designed by the Bangladeshi NGO BRAC, (described in detail below) is currently being evaluated in Honduras, Peru, Karnataka, West Bengal, Bangladesh and Pakistan. Each country has a different research team and a different local implementation partner. Studies of interest rate sensitivity replicating

Karlan and Zinman (2008) are currently under way in Ghana, Peru (in two separate locations with two different partners), Mexico, and Philippines (in three separate locations with two different partners). Microcredit impact evaluations are happening simultaneously in Morocco, urban India, Philippines (in three separate locations), and Mexico. Business training is being evaluated in Peru, the Dominican Republic, urban India, and Mexico. Similar programs to encourage savings are being evaluated in Peru, Philippines, Ghana and Uganda. It thus seems that there is enough interest among funding agencies to fund these experiments, and enough willing researchers to carry them out. For example, in the case of the several on going ultra-poor experiments, the Ford Foundation is funding all of them, in an explicit attempt to gain more understanding of the program by evaluating it in several separate locations. Innovations for Poverty Action (an NGO founded by Dean Karlan), which has been leading the effort for many of these replications is hosting the grant, but the research teams and the implementation partners are different in each country. The different research teams share evaluation strategies and instruments, to make sure that different results represent differences in the contexts, rather than evaluation strategies.

Those studies are still ongoing, and their results will tell us much more about the conditions under which the results from programs are context dependent. Systematic tests on whether the results differ across sites will be needed. One approach will be to treat the different sites as covariates, and use the non-parameteric test proposed by Crump et al (forthcoming) to test whether the effect is different in any of the sites. If heterogeneity is found, a more powerful test would be whether heterogeneity still remains after accounting for the heterogeneity of the covariates. Another way to proceed would be to run the non-parametric regressions proposed by Crump et al (forthcoming) and test whether the treatment effect conditional on the covariates is equal for all the site dummies. While not directly proposed by Crump at al (forthcoming), this would be a straightforward extension. The point is obviously not that every result from experimental research generalizes, but that we have a way of knowing which ones do and which ones do not. If we were prepared to carry out enough experiments in varied enough locations, we could learn as much as we want to know about the distribution of the treatment effects across sites conditional on any given set of covariates.

In contrast, there is no comparable statement that could be made about observational studies. While it may be possible to identify a particular quasi-experiment that convincingly delivers the

correct treatment effect, it seems highly unlikely that such a quasi-experiment could be replicated in as many different settings as one would like. Moreover, with observational studies, one needs to assume non-confoundedness (i.e. that the identification assumptions are valid) of all the studies to be able to compare them. If several observational studies give different results, one possible explanation is that one or several of them are biased (this is the principle behind an over-identification test), and another one is that the treatment effects are indeed different.

However it is often claimed---See Rodrik (2008) for example---that environmental dependence is less of an issue for observational studies because these studies cover much larger areas and as a result, the treatment" effect is an average across a large number of settings and therefore more generalizable.[4] In this sense, it is suggested, there is a tradeoff between the more "internally" valid randomized studies and the more "externally" valid observational studies.

However this is actually not necessarily true. A part of the problem comes down to what it means to be generalizable: it means that if you take the same action in a different location you would get the same result. But what action and what result? In cross-area studies which compare, say, different types of investments, the fact that the action was the same and that the results were measured in the same way must be taken on faith, a decision to trust the judgment of those who constructed the data set and pooled a number of programs together under one general heading. For example, "education investment" could mean a number of different things. The generalizable conclusion from the study is therefore at best the impact of the average of set of things that happened to have been pooled together when constructing the aggregate data.

There is also a more subtle issue about generalizations, which arises even when we evaluate very well defined individual programs. The fact that a program evaluation uses data from a large area, does not necessarily mean that the estimate of the program effect that we get from that evaluation is an average of the program effects on all the different types of people living in that large area (or all the people who are plausible program participants). The way we estimate the program effect in such cases is to first try to control for any observable differences between those covered by the program and those not covered (for example using some kind of matching) and then looking at how those in the program perform relative to those who are not.

[4] Note that not all randomized experiments are small scale. For example, the mandated representation programs we mentioned above was actually implemented nationwide in India. While Chattopdhyay and Duflo (2004) originally looked at only two (very different) States, Topalova and Duflo (2004) extend the analysis to all the major Indian States.

But it is possible that once we match like with like, either almost everyone who is in a particular matched group is a program participant or everyone is a non-participant. There are several methods to deal with this lack of overlap between the distribution of participants and non participants (Heckman, Ichimura and Todd, 1997; Heckman, Ichimura, Smith and Todd, 1998; Rubin 2006—see a review in Imbens and Wooldridge, 2008), but in all cases, the estimate will be entirely driven by the sub-groups in the population where, even after matching, there are both enough of participants and non-participants, and these sub-groups could be entirely non-representative. And while we can identify the observable characteristics of the population driving the estimate of the treatment effect (though this is rarely done) we have no way of knowing how they compare to the rest of the population in terms of un-observables. In the words of Imbens and Wooldridge, "a potential feature of all these methods [that improve overlap between participants and non-participants] is that they change what is being estimated (…) This results in reduced external validity, but it is likely to improve internal validity". Thus, the trade-off between internal and external validity is present as well in observational studies. By contrast as long as the compliance rates among those chosen for treatment in an experiment is high, we know that the affected population is at least representative of the population chosen for the experiment. As is well known (see Imbens and Angrist, 1994), the same point also applies to instrumental variables estimates: the "compliers" in an IV strategy, for whom the program effect is identified, may be a small and unrepresentative subset of the population of interest.

The point made by Heckman (1992) still remains. If randomized evaluations can only be carried out in very specific locations or with specific partners, precisely because they are randomized and not every partner agrees to the randomization, replication in many sites does not get rid of this problem. This is a serious objection (closely related to the compliance problem that we discuss below: it is compliance at the level of the organization), and one that is difficult to refute, since no amount of data could completely reassure us that this is not an issue. Our experience is that, in the context of developing countries, this is becoming less and less of an issue as randomized evaluations gain wider acceptability: evaluation projects have been completed with international NGOs, local governments, and an array of local NGOs. This will only improve if randomized evaluation comes to be recommended by most donors, as it will mean that the willingness to comply with randomization does not set organizations apart any more.

A more serious issue in our experience is the related fact that what distinguishes possible partners for randomized evaluations is competence and a willingness to implement projects as planned. These may be lost when the project scales up. It is important to recognize this limit

when interpreting results from evaluations: finding that a particular program, when implemented somewhere, has a given mean effect leaves open the problem how to scale it up. Not enough effort has taken place so far in trying “medium scale” evaluation of programs that have been successful on a small scale, where these implementation issues would become evident.

That said, this problem is not entirely absent for observational studies either, especially in developing countries. Not all programs can be convincingly evaluated with a matching study. Large data sets are often required (especially if one wants to improve external validity by focusing on a large area). In some cases, data is collected on purpose for the evaluation, often with the assistance of the country statistical office. In this case, the country needs to accept the evaluation of a large program. Large programs are politically more sensitive to evaluate than pilot programs, since they are usually well publicized, and so countries may be strategic with respect to the choice of programs to evaluate. In other cases, regular large scale surveys (such as the NSS survey in India, the Susenas in Indonesia, etc.) can be used. But not all developing countries have them (though data sets like the DHS, which are available for many countries, have certainly ameliorated the issue). There thus is also a potential bias (although quite different from that of randomized evaluation) in the types of countries and programs that can be evaluated with observational data.

The point is not that generalizability is not an issue for the experimental/quasi-experimental approach, but it is not obviously less of an issue for any other approach.

### 2.2 Compliance issues

We already made the point that a high compliance rate makes it easier to interpret the treatment effect, and generalize the results. The experiments in development economics have often been carried out by randomizing over a set of locations or cluster (villages, neighborhoods, schools) where the implementing organization is relatively confident of being able to implement. At the level of the location the take up rate is therefore relatively high, often 100%. It should be emphasized that this just means that the treated sample is likely to be a random sub-set of the set of locations that were selected for the program. The actual individuals who benefited from the treatment are of course not guaranteed to be a random subset of the population of those locations, but it is assumed that the selection at this level mirrors the selection that an actual program (i.e. not just under experimental conditions) would induce, and that the treatment on

the treated parameter of an IV estimate using “treatment” village as the instrument is the parameter of interest.

Heckman (1992) was specifically concerned with the interpretation of randomized experiments in the US where individuals were offered the option to participate in a job training program. Take up was low and potentially highly selected, which would be fine if we wanted to know the effect of offering people such an option, but not if the plan was to make it compulsory, for example, for all welfare recipients. Similar issues also arise in some of the developing country experiments. For example, Karlan and Zinman’s (2007) study of the effect of access to consumer credit starts from a population of those whose loan application was rejected by the bank. Then they asked the loan officers to identify a class of marginal rejects from this population and randomly “un-rejected” a group of them. However the loan officers still had discretion and used it to reject about half of those who were un-rejected. The experiment identifies the effect of this extra credit on the population of those who remained “un-rejected”: It appears to have raised the likelihood that the person remains employed as well as their incomes. However while it provides (very valuable) evidence that consumer credit might be good for some people, given the unusual nature of the treated population (the twice un-rejected) there is some concern about what to make of the actual magnitude of the effect.

Another point made by Heckman is that randomized evaluations are not the best method to study who takes up programs once they are offered to them and why. This is not necessarily the case, as randomization can be used precisely to learn about selection issues: As we discussed above, there are now several studies where the randomization is specifically designed to measure the selection effect, which would be very difficult to do in any other way (Karlan and Zinman, 2005; Ashraf, Berry and Shapiro, 2007; Cohen and Dupas, 2007) discussed above. To learn more about selection, Cohen and Dupas (2007) collected hemoglobin level of women who purchased bed net at different prices. They were interested in whether women who purchase nets only when they free are less likely to be anemic. In other studies, although the evaluation is not specifically designed to capture selection effect, the take up among those offered the program is of special interest, and baseline data is specifically collected to this effect. For example, in Ashraf, Karlan and Yin (2006), one important outcome of interest is who takes up a self-control device that helps people save.

In yet other cases take up is not an issue, because the treatment is in the nature of a pure gift, unlike the offer of training, which is worthless unless someone is also prepared to put in the time. For example, De Mel, Mckenzie and Woodruff (forthcoming) study the effect of offering each firm in their sample in Sri Lanka about $200 in the form additional capital. They find a large impact on the revenue of the firm, equivalent to a 5-7% return on capital. Cull, Mckenzie and Woodruff (forthcoming) repeat the same experiment in Mexico, and find even larger returns (20-35%) In both these cases the fact that the target firms were very small indeed was crucial: this is what made sure that almost everyone was interested in participating in the program (even if it was a gift, there is always some cost of participation), and allowed such a small gift (which is all that they could afford) to have a discernable impact.

However sometimes even a gift may be refused, as Banerjee, Chattopadhyay, Duflo and Shapiro, who are working with the MFI Bandhan to evaluate their programs to help the ultra poor (one of the several evaluations of this program that we mentioned earlier) discovered, to their surprise. Under this program, villagers who are too poor to be brought into the microfinance net are identified through participatory resource assessments (PRA) and other follow up investigations and then offered an asset (usually a pair of cows, a few goats, or some other productive asset) worth between $25 and $100 with no legal strings attached (though there is expectation that they will take care of it and some follow up), as well as a weekly allowance, and some training. The goal is to see if access to the asset creates a long-term improvement in their standards of living (or do they simply sell the asset and run through the proceeds quickly). The evaluation design assumed that everyone who is offered the asset will grab it, which turned out not to be the case. A significant fraction of the clients (18%) refused the offer: Some were suspicious, because they thought it was part of an attempt to convert them to Christianity; others thought it was a trick to get them into a debt trap—that eventually they would be required to pay it back; other did not doubt the motives of Bandhan, but they did not feel capable of doing a good job taking care of the asset and did not want to feel embarrassed in the village if they lost it.

### 2.3 Randomization issues

The Bandhan example reinforces a point also made in Heckman (1992): that the fact that there is an experiment going on might generate selection effects that would not arise in non-experimental settings. This is one example of a Hawthorne of John Henry effects: the fact of

being part of an experiment (and being monitored) influence its participants. The fact that these villagers were not used to a private organization going around giving away assets was clearly a part of the reason why the problem occurred. On the other hand, Bandhan may not have put in the kind of public relations effort to inform the villagers about why it was being done, precisely because they were not planning to serve the entire population of the very poor in each village.

Most experiments however are careful to avoid this issue. Randomization that takes place at the level of location can piggy-back on the expansion of the organization's involvement in these areas limited by budget and administrative capacity, which is precisely why they agree to randomize. Limited government budgets and diverse actions by many small NGOs mean that villages or schools in most developing countries are used to the fact that some areas get some programs and others do not and when a NGO only serve some villages, they see it as a part of the organization's overall strategy. When the control areas given the explanation that the program had only enough budget for a certain number of schools, they typically agree that a lottery was a fair way to allocate it---they are often used to such arbitrariness and so randomization appear transparent and legitimate.

One issue with the explicit acknowledgement of randomization as a fair way to allocate the program is that implementers will find that the easiest way to present it to the community is to say that an expansion of the program is planned for the control areas in the future (especially when it is indeed the case, as in phased-in design). This may cause problems if the anticipation of treatment leads individuals to change their behavior. This criticism was made in the case of the PROGRESA programs, where control villages knew that they were going to eventually be covered by the program.

When it is necessary for the evaluation that individuals not be aware that they are excluded from the program for the evaluation's sake, ethics committees typically grant an exemption from full disclosure until the end-line survey is completed, at least when the fact of being studied in the control group does not present any risk to the subject. In these cases, participants at the ground level are not told that there is an actual randomization involved. This happens more often when randomization takes place at the individual level (though some individual level randomizations are carried by public lottery). In this case, it is simply revealed to the selected beneficiaries that, for example, they got a loan that they had applied for (Karlan and Zinman, 2007), or that the bank had decided that the interest rate could be lower (Karlan and Zinman, 2005).

### 2.4 Equilibrium effects

A related issue is what is usually, slightly confusingly, called general equilibrium effects (and we prefer to call equilibrium effects, since general equilibrium is essentially a multi-market concept). Program effects found in a small study may not generalize when the program is scaled up nationwide (Heckman, Lochner and Taber 1999). Consider for example what would happen if we try to scale up a program that shows, in a small-scale experimental implementation, that economically disadvantaged girls who get vouchers to go to private schools end up with a better education and higher incomes. When we scale up the program to the national level, two challenges arise: one is that there will be crowding in the private schools (and potentially a collapse of public schools) and the other is that the returns to education will fall because of increased supply. For both reasons the experimental evidence could overstate the returns to the vouchers program.

This phenomenon of equilibrium effects poses a problem that has no perfect solution. However there are clearly many instances where we don't expect to face it: if we want to know which strategy for promoting immunization take up (reliable delivery or reliable delivery plus a small incentive for the mother to remember to immunize her child on schedule) is more cost effective in raising immunization rates and by how much (as in Banerjee, Duflo, Glennerster and Kothari (2008) for example) the experimental method poses no problem. The fact that immunizing the entire district would not require that many more extra nurses helps us here because we can assume that the price of nurses would not go up by very much, if at all. On the other hand, while it is useful to learn that those who got vouchers in Colombia do better in terms of both educational and life outcomes (see Angrist, Bettinger, Bloom and Kremer, 2002; Angrist, Bettinger and Kremer, 2006), it is hard to not worry about the fact that an increase in the overall supply of skills brought about by the expansion of the vouchers program will lower the price of skills. After all, that is precisely one of the reasons why the government might want to carry out such a program.

Equilibrium effects offer the one clear reason to favor large studies over small ones. That does not necessarily mean cross-country style regressions--which often conflate too many different sources of variation to be useful in making causal claims (Acemoglu and Johnson, 2007) on the impact of curing diseases on countries economic growth is an excellent exception) –but rather micro studies using large-scale policy shifts. These are typically not randomized, but often still

offer the opportunity to be careful about causality issues and at the same time, help us with respect to equilibrium effects because many of the equilibrium effects get internalized. A good example of this kind of research is the work of Hsieh and Urquiola (2006) who use a quasi-experimental design to argue that a Chilean school voucher program did not lead to an overall improvement in the skill supply, though it changed sorting patterns across schools. Other studies specifically designed to evaluate potential market equilibrium effects of policies include Acemoglu and Angrist (2000) and Duflo (2004).

Clearly the opportunity to do good quality quasi-experimental studies is not always available and in any case it is likely worth checking if the results are consistent with experimental evidence. For example in the case of vouchers we expect the equilibrium effects to dampen the supply response and therefore expect larger quasi-experimental studies to generate smaller effects than experiments. If we find the opposite, we might start worrying about whether the larger study is reliable or representative. In this sense experiments and non-experimental studies may be complements rather than substitutes.

Another approach is to try to directly estimate the size of the equilibrium effect using the experimental method. In ongoing research, Kremer and Muralidharan study the effect of a vouchers program using a double randomization: they randomize villages where the vouchers are given out as well as who gets vouchers within a village. By comparing the estimates that they will get from the two treatments they hope to infer the size of the equilibrium effect. Of course, this only deals with one level of equilibriation—people can move to the village from outside and leave the village to find work; in this case it may work better to estimate what is happening to the supply of education than to the price of skills—but it is clearly an important start.

A related approach is to combine the results from different experiments: Use one experiment (or more plausibly, quasi-experiment) to try to estimate the elasticity of demand for skills another to estimate the supply of quality teaching and another to estimate how much vouchers contribute to skill-building.  This is a style of work that requires taking a more structural approach since we need to identify what the relevant parameters are. As we discuss in the next sub-section, this kind of work is now beginning to happen but there is clearly a long way to go.

### 2.5 Heterogeneity in Treatment Effects

Most evaluations of social programs focus exclusively on the mean impact. In fact, one of the claimed advantages of experimental results is their simplicity: they are easy to interpret since all you need to do is compare means and this might encourage policymakers to take the results more seriously (see e.g. Duflo, 2004; Duflo and Kremer, 2004). However, as Heckman, Smith and Clements (1997) point out, the mean treatment effect may not be what the policymaker wants to know: Exclusive focus on the mean is only valid under rather specific assumptions about the form of the social welfare function. Moreover from the point of view of the overall intellectual project it clearly makes no sense to restrict the analysis to the naïve comparison of means.

Unfortunately, it turns out that the mean treatment effect (or the treatment effect conditional on covariates) is also the only conventional statistic of the distribution of treatment effects that is straightforward to estimate from a randomized experiment without making any additional assumptions. Of course, in principle, one could compare the entire distribution of outcomes in treatment with that in control: there are tests for the equality of distributions, as well for stochastic dominance (see Abadie, 2002). For example, Banerjee, Cole, Duflo and Linden (2007) show that the distribution of test scores among the students who study in schools that received a Balsakhi first order stochastically dominates that of the treatment group, and most of the gains are seen at the bottom. This is important, since in the program classrooms the children at the bottom were pulled out and given remedial teaching, while those top remain in the classroom. We would therefore expect very different effects on the two groups, and it would bee hard to justify the program if it only helps those at the top. Duflo, Hanna and Ryan (2007) also look at how the camera-based teacher incentive program discussed earlier affects the entire distribution of absence among teachers, and find first order stochastic dominance. However, comparing these distributions does not inform us about the distribution of the treatment effect per se (since the differences in quantiles of a distribution is not the quantile of the difference).

In their excellent review of the recent econometric literature on program evaluation (including the technical details behind much of the material covered here), Imbens and Woolridge (2008) make the case that the distribution of the outcome in treatment and in control (which is always knowable) is all that we could possibly want to know about the program, because any social welfare function should be defined by the distribution of outcomes (or by the distribution of outcomes, conditional on observable variables).

However it is not clear that this is entirely correct. To see the issue in its starkest form, consider the following example. There is a population of 3 people, and we know their potential outcomes if treated and if non treated. Mr 1's potential outcome if non treated is 1, Mr 2's is 2, and Mr 3's is 3. Mr. 1's potential outcome if treated is2, Mr. 2's outcome is 3 and Mr.3 's outcome is negative 4. What should we think of this program? Clearly both in terms of the mean treatment effect and in terms of the overall distribution, the treatment failed: the distribution 1,2,3 of the potential outcome "non-treated" first order dominates the distribution -4, 2,3 of the potential outcome "treated". Should we therefore conclude that a policymaker should always favor control over treatment? Not necessarily, since the treatment makes a majority better off and the policymaker might care about the "greatest good of the greatest number". And even if we disagree with the policymaker's preferences here, it is hard to argue that the evaluator should dictate the choice of the objective function.

Once we recognize that we might care about identifying the set of people (from an ex ante undifferentiated group) who moved up or down due to the treatment, there is obviously a problem. There is no way to extract this information from the distribution of outcomes in treatment and in control, a fact that is closely related to Heckman's (1992) observation that even experiments do not deliver quantile treatment effects without additional assumptions.

This is of course a logical problem, and not a problem with experiments *per se* or any other specific estimation strategy—the relevant information is simply not there. In the setting of a randomized social experiment Heckman, Smith and Clements (1997), show that by introducing additional behavioral assumptions (in effect, modeling the decision to participate as a function of the potential outcomes under treatment and non-treatment) allows estimation of rather precise bounds on features of the distribution of the treatment effect. These techniques do also apply in non-experimental settings, but the authors point out that they may be particularly handy with experimental data both because one "can abstract from selection problems that plague non-experimental data", and because the experimental setting guarantees that there is balance in the support of the observable variables, which is something they rely on.

Our view is that experimental research would gain something by engaging more with this body of research. Reporting some more "assumption-dependent" results along with the more "assumption-free" results that are usually reported in experiments (and making the necessary *caveat emptor*) can only enrich experimental work. However, experiments still have the advantage over methods that, with very few assumptions, one can know very important aspects

of impact of the treatment (such as the mean for any subgroup). The fact that we may want to go beyond these measures, and to do so we might need to invoke assumptions that might make random assignment less important cannot possibly be counted in favor of methods not based on radom assignment

Moreover, a lot of the heterogeneity that features prominently in people's objective functions (as against a lot of heterogeneity that drives economic outcomes) is not really about unobserved differences in people's characteristics, but about potentially observable differences. For example, in the balsakhi experiment (Banerjee, Cole, Duflo and Linden, 2007), we not only observed that the distribution of test scores in treatment first order stochastically dominated that in control; we also saw that those who had low baseline scores gained the most. From the point of view of the implementing organization, Pratham, this was what really mattered, but we could only know this because we had baseline test scores. In other words, we need to start the experiment with clear hypotheses about how treatment effects vary based on covariates, and collect the relevant baseline data.

Fortunately, recent econometric research can help us quite a lot here. Crump et al. (forthcoming), already discussed above, develop two non-parametric tests of whether there is heterogeneity in treatment effects: one for whether the treatment effect is zero for any sub-population (defined by covariates), and one for whether the treatment effect is the same for all sub-populations (defined by covariates).

In addition, treatment effects can be estimated for different sub-groups. One difficulty here is that if the sub-groups are determined ex post, there is a danger of "specification searching", where researchers and policymakers ex post choices to emphasize the program impact on one particular sub-group. Here again, as in the Heckman, Smith and Clements (1997) application, theory can help by telling us what to expect. Specifying ex ante the outcomes to be looked at and what we expect from them (as is encouraged in the medical literature) is another possibility. Of course we might still want to try to learn from possibly interesting (but ex ante unexpected) differences in the treatment effect. This is another place where replication can help: when a second experiment is run, it can be explicitly set up to test this newly generated hypothesis. For example, Karlan and Zinman (2007) find very different results for men and women---men are subjected to moral hazard but not much adverse selection while women are the reverse. These differences were not expected, and it is hard to know what to make of them. But once the study is replicated elsewhere, these can form the basis of a new set of hypotheses to be tested (see

Duflo, Kremer and Glennerster, 2008, for a more detailed discussion of these and other design issues).

Finally, a recent literature (Manski 2000, 2002, 2004, Deheijia, 2005, Hirano and Porter, 2005) seeks to make all this less ad hoc. They want to integrate the process of evaluation and learning into an explicit framework of program design. They therefore try to put themselves explicitly in the shoes of the policymaker trying to decide whether or not to implement a program, but also *how* to implement it (should the program be compulsory? Should the administrator be given some leeway on who should participate?). They allow the policymaker to be concerned not necessarily only with expected income gain, but with expected utility gain (taking into account risk aversion), and hence with potential increase or decrease in the variability of the outcome with the treatment status. The policymaker has access to covariates about potential beneficiaries as well as to the results from randomized experiments. This literature tries to develop a theory of how the administrator should decide, taking into account both heterogeneity and uncertainty in program benefits conditional on covariates. As far as we know, these tools have not been used in development economic research. This is a fruitful avenue for future work.

### 2.6 Relationship with Structural Estimation

Most of the early experimental literature focused on reduced form estimates of the program effect. But there is no reason not to also use that data to extract structural parameters wherever possible. While this will require us to make more assumptions, the structural estimates can be used to cross-check the reduced-form results (are the results reasonable if they imply an elasticity of labor supply of $x$ or an expected return on schooling of $y$? ) and more generally to bolster their external validity. Moreover, if we are comfortable with the assumptions underlying the estimates, it is possible to derive policy conclusions from them that go well beyond what could get from the reduced form.

Early examples of this method include Attanasio and Meghir and Santiago (2002) and Todd and Wolpin (2006), which both use PROGRESA data. Attanasio, Meghir and Santiago are interested in evaluating the program impact, while allowing, for example for anticipation effects in the control (which cannot be done without making some additional assumptions). They find no evidence of anticipation effects. Todd and Wolpin (2006) want to use the experiment as a way

to validate the structural model: they estimate a structural model outside the treated sample, and check that the model correctly predicts the impact of the treatment. Another example of the potential of marrying experiments and structural estimation is Duflo, Hanna and Ryan (2007). After reporting the reduced form results, the paper exploits the non-linear aspect of Seva Mandir teacher incentive schemes (teachers received a minimum wage of $10 if they were present less than 10 days in the month, and a bonus of $1 for any extra day above that) to estimate the value of the teacher of not going to school and the elasticity of their response with respect to the bonus. The model is extremely simple (by coming to school in the early days of the months, the teacher is building up the option of getting $1 extra dollar a day by the end, and giving up a stochastic outside option of not going this day), but gives rise to interesting estimation problems, once we want to introduce heterogeneity and serial correlation in the shock received by the teacher on the outside option in a realistic way. Like Todd and Wolpin, this paper then compares the predictions of various models to both the control, and to a “natural experiment” where Seva Mandir changed their payment rules (after the experiment period was over). This exercise showed that accounting for heterogeneity and serial correlation is important, since only those simulations come close to replicating the control group means and the distribution of absence under the new rules.

In principle it ought to be possible to go even further in exploiting the complementarity between structural estimation and experiments. As mentioned already, one advantage of experiments s that is the flexibility with respect to data collection and the choice of treatments: It should be possible to design the experiment to facilitate structural estimation by making sure that the experiment includes sources of variation that would help identify the necessary parameters and collecting the right kind of data. One could also estimate a structural model from baseline data before the experimental results are known, in order to perform a “blind” validation of the structural models. However we are yet to see examples of this kind of work: the examples we discussed exploited ex post variation in the way the program was implemented, rather than introducing it on purpose.

### 2.7 Relation to theory

We have already made the case that experiments can be and have been very useful for testing theories (see Banerjee (2005) and Duflo (2006) for a longer treatment of these issues). The fact that the basic experimental results (e.g. the mean treatment effect) do not depend on the theory

for their identification, means that a “clean” test of theory (i.e. a test that does not rely on other theories too) may be possible

One place this has clearly been very useful is in making us rethink some basic elements of demand theory. One consistent finding of a number of independent randomized studies of the demand for what might be called health protection products, is that the price elasticity of demand around zero is huge. Kremer and Miguel (2007) found that raising the price of de-worming drugs from 0 to 30 cents per child In Kenya reduced the fraction of children taking the drug from 75% to 19%. Also in Kenya, Cohen and Dupas (2007) find that raising the price of insecticide treated bednets from 0 to 60 cents reduces the fraction of those who buy the nets by 60 percentage points. Raising the price of water disinfectant from 9 cents to 24 cents reduces the fraction who take up the offer in Zambia by 30 percentage points (Ashraf, Berry and Shapiro, 2007). Similar large responses are also found with small subsidies: In India, Banerjee, Duflo, Glennerster and Kothari (2008) found that offering mothers one kilo of dried beans (worth about 60 cents) for every immunization visit (plus a set of bowls for completing immunization) increases the probability that a child is fully immunized by 20 percentage points. And most remarkably, a reward of 10 cents got 20 percent more people in Malawi to pick up the results of their HIV test (Thornton, forthcoming).

Kremer and Holla (2008) reviewing this evidence (and several papers on education with similar conclusions), conclude that these demand elasticities cannot come from the standard human-capital model of the demand for better health, given the importance of the issue at hand. For example, one can imagine that a conventionally rational economic agent might either decide to get their HIV test (knowing one’s status could prolong one’s life and that of others) or he may decide against getting it (the test may be extremely stressful and shameful). What is more difficult to understand is that so many of them change their mind for a mere 10 cents, about something that had a good chance of entirely transforming their lives, one way or the other.

Kremer and Holla (2008) suggest that this pattern of demand is more consistent with a model where people actually want the product but are procrastinating; it is tempting to delay paying the cost given that the benefits are in the future. On the other hand, if it is true that people really want to buy bednets or know their test result but are perpetually unable to do so, then, given the potential life-saving benefits that these offer, they have to be extraordinarily naïve. However, when it comes to financial products, the (experimental) evidence argues against their being that naïve. Ashraf, Karlan and Yin (2007) find that it is those who show particularly hyperbolic

preferences that are particularly keen to acquire commitment devices to lock in their savings, indicating a high degree of self-awareness. Duflo, Kremer and Robinson (2008) find that farmers in Kenya who complain of not having enough money to buy fertilizer at planting time are willing to commit money at harvest time for fertilizer to be used at planting several months later. Moreover, when given ex ante (before the harvest), the choice about when they should come to sell fertilizer, almost half the farmers do request them to come right after harvest, rather than later when they will need fertilizer, because they know that they will have money at after the harvest. However, they request the fertilizer to be delivered to them right away, suggesting that they have at least enough self-control to keep fertilizer at home and not resell it. This suggests that the theory might go beyond the now-standard invocation of self-control problems as a way of dealing with all anomalies.

Sometimes experiments throw up results that are even more troubling to the existing body of theory (see Duflo (2004) for a longer discussion). One striking example that fits no existing economic theory, is from Bertrand, Karlan, Mullainathan, Shafir and Zinman (2008): They find that seemingly minor manipulations (such as the photograph on a mailer) have effects on take up of loans as large as meaningful changes in interest rates.

In all of this experiments are playing the role traditionally played by lab experirments, perhaps with greater credibility. The goal is better theory. But can theory help us design better experiments and interpret experimental results better for better policy design? One possible direction, discussed above, is to use experimental results to estimate structural models. However we also want theory to play a more mundane but equally important role: We need a framework for interpreting what we find. For example, can we go beyond the observation that different inputs into the educational production function have different productivities? Is there any way to group together the different inputs into broader input categories on a priori grounds, with the presumption that there should be less variation within the category? Or on the outcome side, can we predict which outcomes of the educational system should co-move much more closely than the rest? Or is every experimental result sui generis?

The theory that would be useful for this purpose is unlikely to be particularly sophisticated. Rather, like the famous Mincer model, it would just a convenient way to reduce dimensionality, based on a set of some reasonable premises. Banerjee et al. (2008) attempt to do something

like this for the case of local public action but their effort is at best partially successful. More work along these lines will be vital.

## 3. Conclusion

We thus fully concur with Heckman's (1992) main point: to be interesting, experiments need to be ambitious, and need to be informed by theory. This is also, conveniently, where they are likely to be the most useful for policymakers. Our view is that economists' insights can and should guide policy-making (see also Banerjee, 2002). They are sometimes well placed to propose or identify programs that are likely to make big differences. Perhaps even more importantly, they are often in a position to midwife the process of policy discovery, based on the interplay of theory and experimental research. It is this process of "creative experimentation", where policymakers and researchers work together to think out of the box and learn from successes and failures, that is the most valuable contribution of the recent surge in experimental work in economics.

Bibliography

Abadie, Alberto (2002). "Bootstrap Tests for Distributional Treatment Effects in Instrumental Variables Models," *Journal of the American Statistical Association*, 97(457) 284-292.

Abdul Latif Jameel Poverty Action Lab, "Fighting Poverty: What Works?" Issue One, Fall 2005.

Acemoglu, Daron, and Joshua Angrist (2000). "How Large Are Human-Capital Externalities? Evidence from Compulsory Schooling Laws," NBER Macroeconomics Annual, Volume 15, pp. 9-59.

Acemoglu, Daron, and Simon Johnson (2007). "Disease and Development: the Effect of Life Expectancy on Economic Growth," *Journal of Political Economy*, December, Volume 115, pp. 925-985.

Angrist, Joshua, Eric Bettinger, and Michael Kremer, (2006). “Long-Term Educational Consequences of Secondary School Vouchers: Evidence from Administrative Records in Colombia,” *American Economic Review*. Volume 96(3), pp. 847-862.

Angrist, Joshua, Eric Bettinger, Erik Bloom, Michael Kremer and Elizabeth King (2002). “Vouchers for Private Schooling in Colombia: Evidence from Randomized Natural Experiments,” *The American Economic Review*, December, Volume 92(5), pp.1535-1558.

Angrist, Joshua, D. Lang, and Philip Oreopoulos (forthcoming). “Incentives and Services for College Achievement: Evidence from a Randomized Trial,” *American Economic Journal: Applied Economics.*

Angrist, Joshua and Victor Lavy (1999). “Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement,” *The Quarterly Journal of Economics*, Volume 114(2) pp. 533-575.

Angrist, Joshua. and Victor Lavy (forthcoming). “The Effect of High School Matriculation Awards: Evidence from Group-Level Randomized Trials,” *American Economic Review,* see also NBER Working Paper No. 9389 (2002).

Ashraf, Nava, Dean Karlan, and Wesley Yin (2006). “Tying Odysseus to the Mast: Evidence from a Commitment Savings Product in the Philippines,” *Quarterly Journal of Economics* 121(2), pp. 635-672.

Ashraf, Nava, James Berry and Jesse M. Shapiro (2007). “Can Higher Prices Stimulate Product Use? Evidence from a Field Experiment in Zambia,” NBER Working Paper No. 13247.

Attanasio, Orazio, Costas Meghir and Ana Santiago (2001), “Education Choices in Mexico: Using a Structural Model and a Randomized Experiment to evaluate Progresa,” UCL Mimeo.

Banaji, Mahzarin. (2001). ‘Implicit attitudes can be measured,” In The nature of remembering: Essays in honor of Robert G. Crowder, H. L. Roediger, III, J. S. Nairne, I. Neath, & A. Surprenant editors, Washington, DC: American Psychological Association.

Banerjee, Abhijit (2002). “The Uses of Economic Theory: Against a Purely Positive Interpretation of Theoretical Results,” BREAD Working Paper No. 007.

Banerjee, Abhijit (2005) “‘New Development Economics’ and the Challenge to Theory,” *Economic and Political Weekly*, Vol. 40(40), October 1-7, pp. 4340-4344.

Banerjee, Abhijit, Rukmini Banerji, Esther Duflo, Rachel Glennerster, Stuti Khemani (2008) “Pitfalls of Participatory Programs: Evidence from a Randomized Evaluation in Education in India,” MIMEO, MIT, 2008.

Banerjee, Abhijit, Shawn Cole, Esther Duflo and Leigh Linden (2007). “Remedying Education: Evidence from Two Randomized Experiments in India,” *Quarterly Journal of Economics,* Volume 122(3), pp.1235-1264.

Banerjee, Abhijit., Esther Duflo, Rachel Glennerster and Dhruva Kothari (2008). “Improving Immunization Coverage in Rural India: A Clustered Randomized Controlled Evaluation of Immunization Campaigns with and without Incentives,” MIMEO MIT.

Banerjee, Abhijit, Suraj Jacob and Michael Kremer, with Jenny Lanjouw and Peter Lanjouw (2005). “Moving to Universal Education! Costs and Trade offs,” MIMEO, MIT.

Beaman, Lori, Raghabendra Chattopadhyay, Esther Duflo, Rohini Pande and Petia Topalova (2008). “Powerful Women: Does Exposure Reduce Bias?” BREAD Wworking Paper No. 181, NBER working paper number 14198.

Berry, James (2008). “Rotten Kids or Rotten Parents: Child Motivation and Education Decision in India,” MIMEO, MIT.

Bertrand, Marianne, Dolly Chugh and Sendhil Mullainathan (2005). “Implicit Discrimination,” *American Economic Review*, Volume 95(2), pp. 94-98.

Bertrand, Marianne, Simeon Djankov, Rema Hanna, and Sendhil Mullainathan (forthcoming). "Corruption in Driving Licensing Process in Delhi," *Quarterly Journal of Economics*.

Bjorkman, Martina and Jakob Svensson (2007). "Power to the People: Evidence from a Randomized Field Experiment of a Community-Based Monitoring Project in Uganda," Community-Based Monitoring of Primary Health Care PCEPR Working Paper No. 6344.

Bleakley, Hoyt (2007). "Disease and Development: Evidence from Hookworm Eradication in the American South," *Quarterly Journal of Economics*, Volume 122(1), pp. 73-117. Blundell, Whitney Newey, Torsten. Persson, editors, Cambridge University Press, Vol. 2(42) (see also BREAD Policy Paper No. 002, 2005).

Bobonis, Gustavo, Edward Miguel, Charu Puri Sharma (2006). "Anemia and School Participation," *Journal of Human Resources*, Volume 41 (4), pp.692–721.

Chin, Aimee (2005). "Can Redistributing Teachers Across Schools Raise Educational Attainment? Evidence from Operation Blackboard in India," *Journal of Development Economics* 78, pp. 384-405.

Cull, Robert, David McKenzie and Christopher Woodruff (forthcoming). "Experimental Evidence on Returns to Capital and Access to Finance in Mexico," *World Bank Economic Review*.

Cohen, Jessica and Pascaline Dupas (2007). "Free Distribution or Cost-Sharing? Evidence from a randomized malaria prevention experiment," Brookings Institution Global Working Paper No.14.

Crump, Richard, Joseph Hotz, Guido Imbens, and OscarMitnik, (forthcoming), "Nonparametric Tests for Treatment Effect Heterogeneity," *Review of Economics and Statistics*.

Dehejia, Rajeev (2005). "Program Evaluation as a Decision Problem," *Journal of Econometrics*, Volume 125, pp. 141-173.

Duflo, Esther (2004a). “The Medium Run Consequences of Educational Expansion: Evidence from a Large School Construction Program in Indonesia,” *Journal of Development Economics* Volume 74(1) pp. 163-197.

Duflo, Esther (2004b). “Scaling Up and Evaluation,” in Accelerating Development, Francois Bourguignon and Boris Pleskovic, editors, World Bank and Oxford University Press: Washington, DC and Oxford, 2004, pp. 342-367.

Duflo, Esther (forthcomoing). “Field Experiments in Development Economics,” forthcoming in *Advances in Economic Theory and Econometrics*, Richard Blundell, Whitney Newey, Torsten. Persson, editiors, Cambridge University Press, Volume 2(42), see also BREAD Policy Paper No. 002, 2005.

Duflo, Esther and Raghabendra Chattopadhyay (2004). “Women as Policy Makers: Evidence from a Randomized Policy Experiment in India,” *Econometrica,* Volume 72(5), pp.1409-1443.

Duflo, Esther, Pascaline Dupas and Michael Kremer (2008). “Peer Effects, Pupil Teacher Ratios, and Teacher Incentives: Evidence from a Randomized Evaluation in Kenya,” MIMEO, MIT.

Duflo, Esther, Pascaline Dupas, Michael Kremer, and Sameul Sinei (2006). “Education and HIV/AIDS Prevention: Evidence from a randomized evaluation in Western Kenya,” World Bank Policy Research Working Paper No.402.

Duflo, Esther, Rema Hanna, and Stephen Ryan (2007) “Monitoring Works: Getting Teachers to Come to School,” NBER Working Paper No. 11880, 2005; BREAD Working Paper No. 103.

Duflo, Esther and Michael Kremer (2004). “Use of Randomization in the Evaluation of Development Effectiveness,” in Evaluating Development Effectiveness (World Bank Series on Evaluation and Development, Volume 7, Osvaldo Feinstein, Gregory K. Ingram and George K. Pitman, editors, Transaction Publishers: New Brunswick, NJ, 2004, pp. 205-232.

Duflo, Esther, Michael Kremer, and Rachel Glennerster “Using Randomization in Development Economics Research: A Toolkit,” in Handbook of Development Economics. Elsevier-North Holland John Strauss and Paul Schultz, editors, Volume 4.

Duflo, Esther, Michael Kremer and Jonathan Robinson (2008a). “How High are Rates of Return to Fertilizer? Evidence from Field Experiments in Kenya,” *American Economic Review Papers and Proceedings*, Volume 98(2), pp. 482-488.

Duflo, Esther, Michael Kremer and Jonathan Robinson (2008b). “Why are Farmers not using Fertilizer? Procrastination and Learning in Technology adoption,” MIMEO, MIT.

Dupas, Pascaline (2007). “Relative Risks and the Market for Sex: Teenage Pregnancy, HIV, and Partner Selection in Kenya,” MIMEO, UCLA.

Gine, Xavier, Dean Karlan and Jonathan Zinman (2008). “Put Your Money Where Your Butt Is: A Commitment Savings Account for Smoking Cessation,” MIMEO, Yale University.

Glewwe, Paul, Nauman Ilias, and Michael Kremer (2003). “Teacher Incentives”. MIMEO, Harvard.

Glewwe, Paul, Michael Kremer and Sylvie (forthcoming). “Many Children Left Behind? Textbooks and Test Scores in Kenya,” *American Economic Journal, Applied Economics.*

Glewwe, Paul, Michael Kremer, Sylvie, and E. Zitzewitz (2004). “Retrospective vs. Prospective Analyses of School Inputs: The Case of Flip Charts in Kenya,” *Journal of Development Economic.* Volume 74(1), pp. 251-268.

Heckman, James J. (1992). “Randomization and social policy evaluation,” in

Evaluating Welfare and Training Programs, editors Charles Manski and I. Garfinkel. Cambridge, MA: Harvard University Press. (also available as NBER Technical Working Paper No.107, 1991).

Heckman, James J., Hidehiko Ichimura, J. Smith, and Petra Todd, (1998). “Characterizing Selection Bias Using Experimental Data,” *Econometrica* Volume 66, pp. 1017-1098.

Heckman, James J., Hidehiko Ichimura, and Petra Todd, (1997). “Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Program,” *Review of Economic Studies,* Volume 64, pp.605-654.

Heckman, James J., Lance Lochner, and Christopher Taber, (1999). “Human Capital Formation and General Equilibrium Treatment Effects: A Study of Tax and Tuition Policy,” Fiscal Studies Volume 20(1), pp. 25-40.

Heckman, James J., Jeffrey Smith, and Nancy Clements, (1997). “Making The Most Out Of Programme Evaluations and Social Experiments: Accounting For Heterogeneity in Programme Impacts,” *Review of Economic Studies*, Volume 64, pp.487-535.

Hirano, Keisuke, and Jack Porter (2005). “Asymptotics for Statistical Decision Rules” *Econometrica.* Volume 71(5), pp. 1307-1338.

Holla, Alaka and Kremer, Michael (2008). “Pricing and Access: Lessons from Randomized Evaluation in Education and Health,” MIMEO, Harvard University.

Hsieh, Chang-Tai and Miguel Urquiola (2006). “The Effects of Generalized School Choice on Achievement and Stratification: Evidence from Chile’s Voucher Program,” *Journal of Public Economics.* Volume 90, pp.1477–1503.

Imbens, Guido, and Joshua Angrist (1994). “Identification and Estimation of Local Average Treatment Effects,” *Econometrica*, Volume 61(2), pp. 467-476.

Imbens, Guido and Jeffrey M. Wooldridge (2008). “Recent Developments in the Econometrics of Program Evaluation,” MIMEO, Harvard University (forthcoming in *Journal of Economic Literature*)

Karlan, Dean and Jonathan Zinman (2005). “Observing Unobservables: Identifying Information Asymmetries with a Consumer Credit Field Experiment,” BREAD Working Paper No. 94.

Karlan, Dean (2005a). “Using Experimental Economics to Measure Social Capital and Predict Real Financial Decisions,” *American Economic Review*, Volume 95(5), pp. 1688-1699.

Karlan, Dean, and Jonathan Zinman (2007). “Expanding Credit Access: Using Randomized Supply Decisions to Estimate the Impacts,” MIMEO, Yale University.

Karlan, Dean and Jonathan Zinman (2008). “Credit Elasticities in Less Developed Countries: Implications for Microfinance,” *American Economic Review*, Volume 98(3), pp.1040-1068.

Kremer, Michael, and Edward Miguel (2007). “The Illusion of Sustainability,” *Quarterly Journal of Economics*, Volume 122(3), pp. 1007-1065.

Kremer, Michael, Edward Miguel, and Rebecca Thornton (forthcoming). “Incentives to Learn,” *Review of Economics and Statistics*, see also NBER Working Paper No. 10971 (2007).

Kremer, Michael, Jessica Leino, Edward Miguel, and Alix Zwane. “Spring Cleaning: Rural Water Impacts, Valuation, and Institutions,” MIMEO, Berkeley.

Manski, Charles, (2000). “Identification Problems and Decisions Under Ambiguity: Empirical Analysis of Treatment Response and Normative Analysis of Treatment Choice”. *Journal of Econometrics,* Volume 95, pp. 415-442.

Manski, Charles, (2002). “Treatment Choice Under Ambiguity Induced by Inferential Problems,” *Journal of Statistical Planning and Inference*, Volume 105, pp. 67-82.

Manski, Charles, (2004). “Measuring Expectations,” *Econometrica*, Volume 72(4), pp. 1329-1376.

Manski, Charles, (2004). “Statistical Treatment Rules for Heterogenous Populations,” *Econometrica*, Volume 2(4), pp. 1221-1246.

McKenzie, David, Suresh de Mel, and Christopher Woodruff (forthcoming). “Returns to Capital: Results from a Randomized Experiment,” *Quarterly Journal of Economics*.

Miguel, Edward and Michael Kremer (2004). “Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities,” *Econometrica*, Volume72 (1), pp. 159-217.

Olken, Benjamin (2007). “Monitoring Corruption: Evidence from a Field Experiment in Indonesia,” *Journal of Political Economy,* Volume 115 (2), pp. 200-249.

Olken, Benjamin, and Patrick Barron (2007b). “The Simple Economics of Extortion: Evidence from Trucking in Aceh,” NBER Working Paper No. 13145, BREAD Working Paper No. 151, CEPR Discussion Paper No. 6332.

Rodrik, Dani (2008). “The New Development Economics: We Shall Experiment, But How Shall We Learn?” MIMEO, Harvard University.

Rubin, Donald, (2006). “Matched Sampling for Causal Effects,” Cambridge University Press, Cambridge, UK.

Thornton, Rebecca, (2007). “The Demand for and Impact of HIV Testing: Evidence from a Field Experiment,”forthcoming, *American Economic Review*.

Todd, Petra, and Kenneth I. Wolpin. (2006). “Using Experimental Data to Validate a Dynamic Behavioral Model of Child Schooling: Assessing the Impact of a School Subsidy Program in Mexico,” *American Economic Review*, Volume 96(5), pp. 1384–1417.