Field Experiments in Development Economics[1]

Esther Duflo
Massachusetts Institute of Technology
(Department of Economics and Abdul Latif Jameel Poverty Action Lab)
BREAD, CEPR, NBER

January 2006

Prepared for the World Congress of the Econometric Society

Abstract

There is a long tradition in development economics of collecting original data to test specific hypotheses. Over the last 10 years, this tradition has merged with an expertise in setting up randomized field experiments, resulting in an increasingly large number of studies where an original experiment has been set up to test economic theories and hypotheses. This paper extracts some substantive and methodological lessons from such studies in three domains: incentives, social learning, and time-inconsistent preferences. The paper argues that we need both to continue testing existing theories and to start thinking of how the theories may be adapted to make sense of the field experiment results, many of which are starting to challenge them. This new framework could then guide a new round of experiments.

[1] I would like to thank Richard Blundell, Joshua Angrist, Orazio Attanasio, Abhijit Banerjee, Tim Besley, Michael Kremer, Sendhil Mullainathan and Rohini Pande for comments on this paper and/or having been instrumental in shaping my views on these issues. I thank Neel Mukherjee and Kudzai Takavarasha for carefully reading and editing a previous draft.

There is a long tradition in development economics of collecting original data in order to test a specific economic hypothesis or to study a particular setting or institution. This is perhaps due to a conjunction of the lack of readily available high-quality, large-scale data sets commonly available in industrialized countries and the low cost of data collection in developing countries, though development economists also like to think that it has something to do with the mindset of many of them. Whatever the reason, the ability to let questions determine the data to be obtained, instead of the data determining the questions that can be asked, has been the hallmark of the best work in empirical development economics and has led to work that has no equivalent in other fields, for example, Townsend (1994) and Udry (1994).

Two concurrent developments have taken place over the last 10 years. First, high-quality, large-scale, multipurpose data sets from developing countries have become more readily available. The World Bank's Living Standard Measurement Surveys and the Rand Corporation Family Life Surveys are two examples of high-quality comprehensive data sets available for many countries. The Demographic and Health Surveys have shorter questionnaires but cover a large number of countries and have generally more than one round per country. Some census data from developing countries are now available on the IPUMS web site, and this collection is growing. Finally, statistical agencies in developing countries have started to make their own surveys, in some cases of excellent quality, available to researchers. These data sources and the wealth of natural experiments available in developing countries have opened a goldmine, which researchers and students have enthusiastically started to exploit. Empirical methods developed in other specialties (notably labor and industrial organization) are now used routinely in development economics. The standards for empirical evidence have risen, putting the field on par with other empirical domains. As a result, studies using an original data set to make an interesting observation not supported by a convincing empirical design are no longer as readily accepted.

Nevertheless, development economists continue with the tradition of doing fieldwork to collect original data: the second development over the last 10 years has been the spread of randomized evaluations in development economics. Randomized evaluations measure the impact of an intervention by randomly allocating individuals to a "treatment" group, comprising individuals who receive the program, and a "comparison" group, comprising individuals who do not, at least for some period of time, receive the treatment. The outcomes are then compared across treatment and comparison groups. Here again, cost is an enormous advantage. While the cost of a good randomized policy evaluation in the U.S. easily reaches millions of dollars, both program costs and data collection costs are much lower in developing countries. This has allowed the practice to generalize beyond a few very well-crafted, major projects to a multiplicity of programs, countries, and contexts. In addition, while some of the well-known randomized evaluations are just that—rigorous evaluations of a particular policy intervention[2]—the tradition of posing the question first and then finding the data to answer it has continued with randomized evaluations.

What many development economists now do is to work closely with implementing agencies—NGOs, private companies, or governments—to develop interventions and evaluate them in a randomized setting. The interventions are designed to answer a specific practical problem in a specific context; for example, how to get teachers to come to school more often, how to help farmers to save more, how to convince parents to get their children immunized, how to fight corruption most effectively. What the economists bring to the table, in addition to evaluation expertise, is prior evidence and theories that help them to predict what should work—and how—and what should not. The evaluation of the program then serves to test of these theories: randomized evaluations have become, in effect, field experiments—a new and powerful tool in the arsenal of the economist. In this essay, "field experiment" refers to the implementation and evaluation, by comparing different treatment groups chosen at random, of an intervention or a set of interventions specifically designed to test a hypothesis or a set of hypotheses.

There remains a clear need for better evaluations of different policy options. As Banerjee and He (2003) point out, what is lacking among development practitioners are not ideas, but an idea of whether or not the ideas work. Mullainathan (2005) argues that self-serving bias, which is perhaps particularly pervasive among those who are the most motivated to help the poor, contaminates many evaluations. Randomized design can, to some extent, alleviate this problem. Elsewhere (Duflo, 2004; Duflo and Kremer, 2004), I have advocated the systematic use of randomized evaluations as a way to improve policy effectiveness, and in Duflo, Glennerster, and Kremer (2005), discussed design issues and technical aspects of running and analyzing randomized experiments. The objective of this paper is different: It is to review the use of field experiments as a tool by development economists and to assess the lessons we have learned from them, the challenges field experiments face and those they pose to core theories in economics, and the areas field experiments leave open for research.

[2] For example, the Government of Mexico requested an evaluation of its conditional cash transfer program, PROGRESA/Oportunidades. Then, using a combination of the randomization inherent in the program and assumptions from economic theory, researchers were able to recover parameters of interest (Attanasio, Meghir, and Santiago, 2002; Todd and Wolpin, 2004).

The remainder of the paper proceeds as follows: Section I reviews the substantive conclusions from field experiments in three domains: incentives, social learning, and inconsistent time preferences. Section II extracts methodological lessons from this experience. It argues that we now need both to continue testing existing theory and to start thinking about how the theories may be adapted to make sense of the results from the experiments. Section III concludes.

I. A few things we have learned from experiments

Field experiments have been designed to shed light on core issues in economics, such as the role of incentives or social learning. In recent years, several experiments have tested some of the hypotheses put forward in behavioral economics. A full review is beyond the scope of this paper, but this section reviews what has been learned from field experiments in three domains: incentives, social learning, and inconsistent time preferences.

1) Incentives

The idea that individuals respond to incentives is at the core of much of economics. The poor performance of government workers in developing countries is often attributed to the weak incentives they face and to the fact that incentives that are in place on the books are not implemented. For example, teachers in India can be suspended for not showing up to school, but a survey in rural India showed that this hardly ever happens, despite very high absence rates (Chaudhury et al., 2005).

Accordingly, many experiments have been set up to study how incentives faced by individuals affect their behavior. A fair number of these experiments have been conducted in schools, with incentives provided either to teachers or to students in the form of rewards for improved performance. Answering these questions is important, since this will tell us whether efforts to reform institutions to provide stronger incentives can have a chance to improve performance.

To obtain a sense of the potential impact of providing high-powered incentives to teachers on absence and learning, Duflo and Hanna (2005) evaluated an incentive program that was actually rigorously implemented in the field. Working in conjunction with Seva Mandir, the implementing NGO, they designed and evaluated a simple incentive program that left no space for manipulation and could be strictly implemented.

Seva Mandir runs non-formal, single-teacher primary education centers (NFEs) in tribal villages in rural Udaipur district, a sparsely populated, arid, and hilly region. Tribal villages are remote and difficult to access, which makes it is very difficult for Seva Mandir to regularly monitor the NFEs. Consequently, absence rates are very high despite the organization's policy calling for dismissal in cases of high absence rates: in a study they conducted in 1995, Banerjee et al. (2005) found an absence rate of 40 percent, and at the beginning of their study, in August 2003, Duflo and Hanna (2005) found an absence rate of 44 percent.

Seva Mandir selected 120 schools to participate in the experiment. In 60 randomly selected schools (the "treatment" group), they gave the teacher a camera with a tamper-proof date and time function and instructed him to take a picture of himself and his students every day at opening time and at closing time. Teachers received a bonus as a function of the number of "valid" days they actually came to school. A "valid" day was defined as a day where the opening and closing pictures were separated by at least 5 hours and a minimum number of children were present in both pictures. The bonus was set up in such a way that a teacher's salary could range from 500 rupees to 1,300 rupees, and each additional valid day carried a bonus of 50 rupees (6 US dollars, valued at PPP, or a little over a dollar at the official exchange rate). In the remaining 60 schools (the "comparison" group), teachers were paid 1,000 rupees and they were told (as usual) that they could be dismissed for poor performance. One unannounced visit every month was used to measure teacher absence as well as teachers' activities when in school.

The introduction of the program resulted in an immediate and persistent improvement in teacher attendance. Over 18 months, the absence rate was cut by almost half in the treatment schools, falling from an average of 42 percent in the comparison schools to 22 percent in the treatment schools. The program was effective on two margins: it completely eliminated extremely delinquent behavior (less than 50 percent presence), and it increased the number of teachers with a perfect or very high attendance record (in treatment schools, 36 percent of teachers were present 90 percent of the time or more; in comparison schools, less than 1 percent were present).

The experiment also provided an ideal setting to test the hypothesis of multitasking (Holmstrom and Milgrom, 1991), in which individuals facing high-powered incentive schemes may change their behavior in such a way that the proximate outcome on which the rewards are based increases, but the ultimate outcome in which the principal is interested remains constant or even decreases. In this case, the incentive was explicitly based only on presence, but Seva Mandir was ultimately interested in improving learning. Teachers may have decided to teach less, once in school. In fact, when in school, the teachers were as likely to be teaching in treatment as in comparison schools, and the number of students present was the same. But, since there were fewer absences, treatment schools taught the equivalent of 88 child-days more per month than comparison schools, a one-third increase in the number of child-days, resulting in a 0.17 standard deviation increase in children's tests scores after one year.

Multitasking did happen, however, in an experiment conducted in Kenya where the incentives provided to teachers were based on the test scores of students in their class (Glewwe, Ilias, and Kremer, 2003). International Child Support (ICS) Africa, the implementing NGO, provided prizes to teachers in grades 4 through 8 based on the performance of the school as a whole on the district exams each year. All teachers who taught these grades were eligible for a prize. Prizes were awarded in two categories: "Top-scoring schools" and "Most-improved schools." Schools could not win in more than one category. Improvements were calculated relative to performance in the baseline year. In each category, three first, second, third, and fourth prizes were awarded. Overall, out of the 50 schools participating in the program, 24 received prizes of some type, and teachers in most schools should have felt that they had a chance of winning a prize. Prizes were substantial, ranging in value from 21 percent to 43 percent of typical monthly salaries of teachers. The comparison of the 50 treatment and 50 control schools suggested that this program did improve performance on the district exams (by about 0.14 standard deviations), but it had no effect on teacher attendance. Instead of attending more often, teachers held more test preparation sessions. This, the authors conclude, was rational based on the (limited) evidence on what is most effective in improving test scores over the short horizon. However, these preparation sessions probably cannot be counted as substitutes for the regular classes: for one, they did very little for long-term learning, as evidenced by the fact that once the program ended, those who had been in the program schools did not outperform students in the comparison schools. In this case, we see teachers responding to incentives in the most cost-effective way possible from their point of view.

Combined, the results of the two experiments make sense. Coming to school regularly is probably the most costly activity for a teacher: there is the opportunity cost of attending, the pressure to dispatch other duties in an environment where nobody strongly expects the teachers to show up every day, and the distance to travel. Once they are in school, their marginal cost of teaching is actually fairly low. Thus, it is not surprising that an effective incentive program rewarding presence does not effectively lead to a reduction in the provision of other inputs when in school. An incentive based on test scores, on the other hand, leaves teachers with ample room (and incentive) to manipulate their way around paying the cost of regular attendance: rather than coming more often, they find other ways to improve test scores.

The camera experiment shows that a straightforward incentive program, mechanically implemented in a relatively simple environment (a single-teacher school), is a very effective way to reduce absence in schools. But most school systems, being larger, more complicated, centralized hierarchies, do not implement incentive schemes this directly; instead, they rely on the mediation of people in the hierarchy, such as inspectors and headmasters. This experiment suggests that one of the reasons why the incentives may fail in these systems is not so much that people do not react to incentives but that the mediators pervert the incentive system. Indeed, Kremer and Chen (2001), in an experiment they conducted in Kenya in partnership with ICS Africa, found that when implemented by headmasters, incentives tend to lose their power. ICS Africa introduced an incentive program for pre-primary school teachers, and the headmaster was entrusted with monitoring the presence of the teacher. At the end of the term, a prize (a bicycle) was to be given to teachers with a good attendance record. If a teacher did not have a good attendance record, the money would remain with the school, and could be used as the headmaster and the school committee saw fit. In all treatment schools, the headmasters marked the teachers present a sufficient number of times for them to get the prize. However, when the research team independently verified absence through unannounced visits, they found that the absence rate was actually at the same high level in treatment and in comparison schools. It seems that in order to avoid the unpleasantness of a fight, or out of compassion for the pre-school teachers, or because they wanted to give the impression of running a tight ship (after all, ensuring presence was part of their regular duties), headmasters actually cheated to make sure that pre-school teachers could get the prizes. This suggests that whenever human judgment is involved, in an environment where rules are often bent, incentives may be easily perverted, either, as in this case, in an equitable direction or else to favor some specific individuals or groups.

The results of these experiments all conform to the priors most economists would have, namely, that individuals respond to incentives and will try to pervert the incentives if they can do so at little cost. As such, the findings of the camera experiment may not provide immediate policy levers or options for policy action, since they do not tell the policymaker that self-policing using cameras would be possible in the larger government schools that constitute the more general and larger policy concern. But combined with the findings on incentive schemes mediated by headmasters, they do clarify the policy possibilities, by telling us that the main, and more general, impediment seems to be in the implementation of incentives schemes, rather than in the effectiveness of the approach.[3] Other experiments offering incentives to students to perform well in school (Angrist and Lavy (2002) in Israel; Kremer, Miguel, and Thornton (2004) in Kenya) have also yielded results that accord well with this prior: when provided with rewards to perform well on exams, students increased their performance. The results of the latter experiment, however, are a little difficult to reconcile with the results of the study on incentives for teachers we discussed above (Glewwe, Ilias, and Kremer (2003)). Incentives for students based on test scores led to durable (rather than only temporary) improvement in test scores as well as a reduction in teacher absence, whereas when teachers were rewarded based on test scores, the improvement in test scores was only temporary, and there was no reduction in absenteeism. The argument of multitasking, which made sense in the teacher incentive context, should apply just the same in the student incentive case, which was based on the same type of tests. The multitasking theory alone provides little guidance on why the reaction was different in one context than in the other.

Still, these experiments show that when incentives are expected to matter, they do. But do they matter even when we would not expect them to? Some experiments designed to answer this question have led to surprising results. For instance, Abhijit Banerjee, Esther Duflo, and Rachel Glennerster, in collaboration with Seva Mandir, set up an experiment to test the impact of small incentives for children's immunization. The rate of immunization in the sample at the baseline was abysmally low—only 1 percent of children are fully immunized by the age of 2. This is surprising, given that immunization is provided free of charge at local health centers, and that the individual benefits of immunization for the child are extremely large: it protects the child against deadly diseases that are still prevalent in the area (measles, tuberculosis, tetanus, and so on). But the absence rate at the center is 45 percent (Banerjee, Deaton, and Duflo (2004)). One reason the immunization rate is so low may be that the parents hesitate to travel to the health center because they are not sure that the center will be open. Given the distance, the travel cost, combined with the uncertainty, may be large enough to compensate the existing benefits. To test the hypothesis that it is the travel costs that prevent parents from immunizing their children, regular immunization camps were set up in the 68 randomly selected villages (out of 135 study villages). The camps were held on a fixed day of the month. Villagers were all informed of the camps in a village meeting, and those who had children to be immunized were

[3] Given the weak incentives provided by formal systems, many have been tempted to propose the use of community monitoring as an alternative to external monitoring. Banerjee and Duflo (2005) review evidence from a variety of field experiments on this topic. Existing evidence is not encouraging for the community model: in all the experiments that have been conducted for far (in education as well as other domains, such as corruption in roads –see Olken (2005)), entrusting communities to conduct the monitoring has been ineffective.

reminded of the camp by a health worker the day before it was held. The health worker also reminded them of the importance of immunization. The health worker (usually a man) had an incentive to do his job as well as possible, since he was paid as a function of the number of children immunized. In half of these camps, mothers were additionally given a kilogram of lentils (a value of 20 rupees) for each immunization received by a child under 2. (Children under 5 would still be immunized in the camp, but the mother would not receive an incentive for them). The incentive was small and unremarkable (lentils are a staple of the local diet). If the cost of immunization was the main barrier, immunization rates should immediately increase when the camps are set up, and the incentive should not matter very much. Even though the experiment is still ongoing, the preliminary results very clearly indicate that the story is more complicated. While the camps are well attended, and are indeed associated with an increase in immunization rates, the attendance is more than five times as large in camps where the lentils are distributed. This increase in attendance is due to crowding in: people traveling a fair distance from other villages; and to the increase in the probability of being immunized among children in villages hosting the camps. A survey of immunization rates among children 0 to 2 years old conducted on 30 families randomly selected from among families in all 135 villages showed the following: In control villages, 4 percent of the 0 to 2-year-olds are "on target" (i.e. they have received the recommended immunization for their age). In villages hosting the camps, 22 percent are "on target." And in the villages participating in the incentive scheme, over 40 percent are "on target," even though the camps have often not been in existence long enough for all the 0 to 2-year-olds to have "caught up" if they started late. A small incentive associated with an activity that has very high returns in the future leads to a very large change in behavior. Either people are not aware of the benefits of immunization (even after they are informed in one large meeting, and reminded of it every month), or, more likely, they tend to delay activities that are a little bit unpleasant in the present (they still need to take a few hours off to take the child to the clinic; immunization makes children cry) even if they have very large returns in the future. For a large number of people, the small but immediate reward is sufficient to solve this problem, whereas a direct subsidy of the activity fails to convince them. Similar results have been found in other contexts. Thornton (2005) finds that very small rewards induce large changes in the probability that someone decides to return to a clinic to find out the results of a (free) HIV-AIDS test. This difference in response indicates that even though individuals are responsive to incentives, they give much more weight to short-term gains. This could be due to a lack of information about what the long-term gains really are, resulting in a very flat indifference curve. Another explanation may be that people have either extremely high discount rates, or more likely, discount rates that are inconsistent over time, with a strong preference for the present. Thus, even when an activity is subsidized to the point where the cost of undertaking it cannot be reduced further, the small, short-term cost does prevent people from participating in it unless the cost is balanced by an equally small reward.

What have we learned from experiments on incentives so far? People seem to be extremely responsive to incentives. They seem particularly responsive to incentives that lead to an immediate reward, relative to large gains in the future, suggesting deviation from the exponential utility model. New experiments should be designed to push this point further. Would people be responsive to delayed incentives if they were given more information or if the incentives were made more salient? Or is it the delayed character of the reward that creates this wedge between a small immediate reward and a large gain in the future?

2) Learning and social effects

The extent to which people learn from each other is a central question in development economics. In particular, the diffusion of new technologies through social networks (neighbors, friends, and so on) has been, and continues to be, intensively studied. The impact of learning on technology adoption in agriculture has been studied especially carefully. Besley and Case (1994) showed that in India, adoption of high-yield variety (HYV) seeds by an individual is correlated with adoption among his neighbors. While this could be due to social learning, it could also be the case that common unobservable variables affect adoption of both the neighbors. To address this issue, Foster and Rosenzweig (1995) focus on profitability. During the early years of the Green Revolution, returns from HYV seeds were uncertain and depended on adequate use of fertilizer. The paper shows that in this context, the profitability of HYV seeds increased with past experimentation, by either the farmers or others in the village. Farmers do not fully take this externality into account, and there is therefore underinvestment in the new technology. In this environment, the diffusion of a new technology will be slow if the neighbors' outcomes are not informative about the individual's own conditions. Indeed, Munshi (2004) shows that in India, in adopting HYV rice, a grain characterized by much more varied conditions, farmers displayed much less social learning than with HYV wheat.

But all these results could still be biased in the presence of spatially correlated profitability shocks. Using detailed information about social interactions, Conley and Udry (2003) distinguish geographical neighbors from "information neighbors"—the set of individuals from whom an individual neighbor may learn about agriculture. They show that pineapple farmers in Ghana imitate the choices (of fertilizer quantity) of their information neighbors when these neighbors have a good shock, and move further away from these decisions when they have a bad shock. Conley and Udry try to rule out the fact that this pattern is due to correlated shocks by observing that the choices made on an established crop (maize-cassava intercropping), for which there should be no learning, do not exhibit the same pattern.

All these papers seek to solve what Manski (1993) has called the "reflection problem": outcomes of neighbors may be correlated because they face common (unobserved) shocks, rather than because they imitate each other. This problem can be solved, however, using an experimental design where part of a unit is subject to a program that changes its behavior. The ideal experiment to identify social learning is to exogenously affect the choice of technology of a group of farmers and to follow subsequent adoption by themselves and the members of their network.

Duflo, Kremer, and Robinson (2005) performed such an experiment in Western Kenya, where less than 15 percent of farmers use fertilizer on their maize crop (the main staple) in any given year, despite the official recommendation (based on results from trials on experimental farms), as well as the high returns (estimated to be greater than 100 percent in these farmers' conditions). In each of six successive seasons, they randomly selected a group of farmers (among the parents of children enrolled in several schools) and provided them with fertilizer and hybrid seeds sufficient for small demonstration plots on their farms. Field officers from ICS Africa guided the farmers throughout the trial, which was concluded by a debriefing session. In the next season, the adoption of fertilizer by these farmers was about 10 percent higher than that of farmers in the comparison group. (Over time, the difference between treatment and comparison farmers declined). However, there is no evidence of any diffusion of this new knowledge: people listed by the treatment farmers as people they talk to about agriculture (their "contacts") did not adopt fertilizer any more than the contacts of the comparison group. Note that this is very different from what would be obtained if one simply regressed a farmer's adoption on his contacts' adoption. The difference suggests that the omitted variable bias, which many of the studies quoted above worried about, is indeed serious: a farmer who has one more contact that uses fertilizer is 10 percent more likely to use it himself in a given season (and this coefficient is significant).

To understand the lack of learning revealed by these results, it is necessary to "unpack" various reasons that may prevent farmers from learning from each other. Note that the trials gave the farmers the opportunity to experiment with fertilizer on their own farms, but it also provided them with additional inputs: the fertilizer was applied with the help of an ICS field officer, who also visited the farmers regularly and helped the farmers compute their rate of return and gave information on results obtained by others at the end of the intervention. The neighbors did not get the benefit of this information.

To distinguish the effect of learning by doing from the effect of the additional information provided, two additional experiments were conducted. In the first, designed to evaluate the impact of learning by doing, each farmer was provided with a starter kit consisting of either enough fertilizer or enough fertilizer and hybrid seeds for a 30-square-meter plot. Farmers were instructed that the kit was sufficient for this amount of space, and they were given twine to measure two plots of the relevant size. Beyond this, there was no monitoring of whether or not (and how) the farmers used the starter kit. Starter kits have been used elsewhere: for instance, the Malawian government distributed 2.86 million such packs beginning in 1998. In the ICS program, field staff explained how to use the inputs but did not formally monitor or measure the yields. Relative to the comparison group, farmers who were provided starter kits were 12 percentage points more likely to use either fertilizer or hybrid seeds. Learning by doing alone seems to affect fertilizer use by about as much as learning from an experiment conducted in one's own field.

The other component of the agricultural trial was that the demonstrations were conducted on the farmer's own plot. If different plots have different characteristics (soil quality, slope, and so on), the learning gained on one farm may not be very useful for a neighbor as it is for the farmer himself. To evaluate the impact of this component, ICS randomly selected one of the farmer's "agricultural contacts" for an invitation to take part in the key stages of the trial (notably planting, harvesting, and the discussion of profitability). After one season, adoption was 17.8 percentage points higher in the first group. This suggests that the effect of watching a demonstration on someone else's plot is as large as the effect of experimenting on one's own plot. It is possible to learn from others.

This last result suggests that if farmers talked to each other, they would be able to learn from each other's experience: the shocks to a farmer's plot are not so large that they make learning impossible. This did not seem to happen, which suggests that the remaining explanation is either that farmers do not talk very much to each other about agriculture, or that they do not trust each other (and they trust the outside (and impartial) experimenter more). The former hypothesis was corroborated by interviewing farmers about themselves and their neighbors and contacts on key parameters of agriculture (date of planting, whether or not the neighbor uses fertilizer, whether or not the neighbor participated in an agricultural trial), and cross-examining the answer: most farmers are either unable to answer regarding their friends or neighbors, or give the wrong answer. Farmers do not appear to know much about each other at all.

If there is diffusion in Ghana and India, but not in Kenya, then there may be another type of externality and source of multiple steady states: when there is very little innovation in a sector, there is no news to exchange, and people do not discuss agriculture. As a result, innovation dies out before spreading, and no innovation survives. When there is a lot of innovation, it is more worthwhile to talk to neighbors, and innovations are in turn more likely to survive.

It is worth noting that, given the accumulated evidence about social learning, the "file drawer bias" would have probably led to the burial of the initial results of "no learning" if they had been obtained in a regression, rather than in an experiment. In a non-randomized study, the researchers would probably have concluded they had done something wrong, and there was nothing interesting in these results. Instead, the initial results of "no learning" prompted a series of additional experiments that helped shed light on the finding. The combination of several experiments examining various aspects of the question should, at a minimum, move our priors about the strength of these effects.

An even more surprising result is obtained by Kremer and Miguel (2003). They use a very similar design in the context of a program to fight intestinal worms in Western Kenya. The program was randomly phased in among three groups of schools, where treatment started in different years (Group 1 started the first year, Group 2 the following year, and finally Group 3 the final year). Because the schools were randomly assigned to each group, conditional on the total number of friends a child had, the number of friends she had in "early treatment" (Group 1 or 2) or "late treatment" (Group 3) schools was exogenous. In 2001 (when the program was just starting in the late treatment schools), the researchers conducted a survey on the number of friends that parents and children in the study schools had in various schools. They then regressed the probability of a child taking the treatment on the number of friends that a child (or her parents) had in the early and late treatment schools, after conditioning on total number of friends the child had and the number of friends she had in her own school. Surprisingly, they found that the more friends a child (or her parents) had in the early treatment schools, the *less* likely she was to take the treatment herself: for each additional social link to an early treatment school, a parent's child is 3.1 percentage points less likely to take the drugs. Further, the effect is too large to be explained by the health externalities effect (if my friend takes the treatment, I do not need to take it). Instead, the researchers attribute it to overoptimistic priors of the family regarding the private health benefits of taking the treatment. When a child's friend takes the medication and does not instantly get much better, the parents actually learn that the medication is less effective than they thought.

Importantly, replicating the usual "due diligence" specifications without using the randomization, produces dramatically different results: when a child's take-up is regressed on average take-up in her school, there is a very significant positive relationship. The correlation is stronger with the take-up in the child's own ethnic group (as in Munshi and Myaux (2002), which uses this strategy as a way to attempt to correct for bias), and there is still a positive correlation between take-up among the child's friends and own take-up. The "effect" is also stronger when more of the child's classmates report a positive experience with the medicine (as in Conley and Udry (2003)), though this difference is not significant.

The experiments reviewed in this section address a very important question for development economics, and solve an identification problem that has proved extremely difficult to address with observational data alone. In both cases, the results differ quite markedly from what the observational studies had found, even when using the same "robust" strategies. These two experiments are clearly insufficient to invalidate prior evidence: they took place in a different setting, and they are both consistent with theoretical models where there is social learning in some conditions but not others.[4] The fact that they are able to replicate the results of previous studies in the same population where they find very different experimental results is, however, troubling: here we cannot argue that it is because the context is different that the results are different. The difference has to come entirely from the methodologies and the assumptions, and implies that the identification assumptions made by the other studies would be violated in this context. In order to believe the previous studies, we now need to believe both that the effects are different and that the assumptions that are invalid in one case are valid in the others. While there is a good a priori argument to make about the former, it is much less clear we can convincingly argue the latter: the argument made by Munshi and Myaux (2002) about communication within and across religious groups, for example, could be essentially replicated for school children of different ethnic groups in Kenya. These experiments make both a methodological point and a substantial point: they cast doubt on the previously "accepted" methodologies, and require revisiting the generally agreed upon notions about social learning.

3) Time inconsistent preferences; demand for commitment and savings

[4] Moreover, it is not the case that no experiment finds any trace of social learning: Duflo and Saez (2003), who conducted the first experiment to implement this design, found very strong social effects in the case of 401k participation in the U.S., consistent with previous non-experimental evidence (Duflo and Saez (2002)). In the case of neighborhood effects on teenager schooling and crime in the U.S., however, experimental evidence suggests that they are much smaller (or even perverse as in this case) than non-experimental evidence would suggest (Kling and Liebman (2005)).

Behavioral economists and psychologists have extensively studied the phenomenon of time-inconsistent preferences. In particular, lab experiments and survey questions have all shown that individuals are impatient in the present and patient in the future: for example, many people who would refuse to get $110 tomorrow instead of $100 today are happy to get $110 in 31 day versus $100 in 30 days. This phenomenon (short-run impatience, long-run patience) is often modeled as discount rates that vary with horizon. People have a very high discount rate for short horizons (decisions about now versus the future), but a very low one for distant horizons. This is often called hyperbolic discounting (Strotz 1956, Ainslie 1992, Laibson 1997). Banerjee and Mullainathan (2005) propose a different model of time-inconsistent preferences, where the individuals must resist immediate temptation (for example, a cousin is sick and wants money now, and it is painful to refuse, even though saving for your children's school fees is the rational decision to make in the long run). Their model makes it explicit that time-inconsistent preferences may arise from the dynamic social context in which individuals are plunged. Time-inconsistent preferences constitute one of the key theories of behavioral economics, and Mullainathan (2005) makes a very convincing argument that they are likely to be central to our understanding of many problems in developing countries, ranging from education to savings.

Yet, while the existence of time-inconsistent preferences is well established in the lab, and while there are institutions (such as ROSCAs in developing countries, 401k plans with withdrawal penalties and Christmas clubs in developed countries) which are consistent with such preferences, there was until recently a dearth of direct evidence of their practical relevance. A key question when analyzing the consequences of these preferences is whether people are sophisticated or naïve in how they deal with their temporal inconsistency. Sophisticated people would recognize the inconsistency and (recursively) form dynamically consistent plans. In other words, they would only make plans that they would follow through on. Naïve people, on the other hand, would not recognize the problem and would make plans assuming that they'd stick with them, only to abandon them when the time comes. Sophisticated hyperbolic discounters will therefore have a demand for commitment devices, whereas naïve hyperbolic discounters will not.

Two recent experimental studies directly test this prediction. Ashraf, Karlan, and Yin (2006) designed a commitment savings product for a small rural bank in the Philippines. Individuals could restrict access to the funds they deposited in the accounts until either a given maturity or a given amount of money was achieved. Relative to standard accounts, the accounts carried no advantage other than this feature. The product was offered to a randomly selected half of 1,700 former clients of the bank.

The other half of the individuals were assigned either to a pure control group or to a group who was visited and given a speech reminding them of the importance of savings.

Relative to the control group, those in the treatment group were 28 percentage points more likely to open a savings account after six months, and their savings increased by 47 percentage points more. The effects were even larger after one year. Prior to offering the products, the experimenters asked the standard hypothetical preference-reversal questions. They found that among women, those who had more tendency to exhibit preference-reversal were also the most likely to take up the product. This study leaves some important points somewhat unresolved: First, the effect of the "marketing" treatment (where the clients were just reminded of the importance of savings and could open a regular account) is also positive, quite large, and it is not possible to statistically distinguish the effect of the commitment savings treatment from the effect of the marketing treatment. Second, the fact that the time-reversal questions predict seed take-up for women, but not for men, is not something that was predicted by the theory.

However, this is one of the few studies that link "lab" evidence to real behavior (other studies in a development context include Binswanger (1980)—followed by many similar exercises in other contexts—and Karlan (2005)). Moreover, this is the first study where the real-life outcome that was studied (the take-up of a commitment savings product) was studied in the context of a randomized experiment. In addition to the substantive points (individuals do take up commitment savings products when they are offered to them, and do set up meaningful targets; they save at least as much with those as with a regular reminder, and probably more), the study makes a methodological contribution regarding the usefulness of time-reversal questions. While the conclusion of the authors is that the results do reflect time-inconsistency in preferences, the results for men suggest that they need more probing before they can be widely used in this way.

Duflo, Kremer, and Robinson (2005), as part of the project discussed earlier, also set up experiments to test whether there is a demand for commitment savings for fertilizer use for maize crop in Western Kenya. They observed that many farmers *plan* to use fertilizer in the next season, or later in the season, but very few end up doing it. The reason they give most frequently is that they have no money when the time come to buy it. Maize crop in Kenya is harvested in two seasons, long and short rain. Fertilizer is administered either at planting (a few weeks after harvest) or at top dressing (a few weeks later, when the maize is knee-high). A farmer needs to save enough between harvest and planting or top dressing to be able to use fertilizer. Over several seasons the researchers worked with ICS to develop, refine, and test in randomized settings a commitment product (called SAFI—Saving and Fertilizer Initiative—*safi* means "pure" in Swahili) for farmers. Each season, a farmer is visited at harvest time, and is given the offer to buy a voucher for fertilizer, valid for whenever he or she wants to redeem it. The voucher can be paid in cash or in maize. The maize is bought at harvest price (a low price), and the fertilizer is sold without a discount. The only advantage to the farmer, in addition to the value of deciding now to use fertilizer later, is that the fertilizer is delivered to his farm, and that the maize is also purchased at his farm.

The final experiment (conducted in 2004) followed a design that allows the testing of several hypotheses. The SAFI program was offered to a group of 420 farmers, randomly selected. In addition, 293 farmers were visited at fertilizer application time. Half of them were offered exactly the same deal. Half of them were offered a 50 percent discount on the fertilizer. Comparing the take-up of the offer to purchase fertilizer in the "visit" group to that in the "SAFI" group allows the researchers to test whether the timing of the decision to purchase the fertilizer is important (as opposed to the free delivery and the ability to sell maize). Comparing the take-up of the offer in the SAFI group to that in the subsidy group lets the researchers benchmark the value of the early purchase relative to a subsidy. In the SAFI group, 40 percent of the farmers bought a voucher for fertilizer. In the subsidy group, 45 percent did. In the visit group, 21 percent did. The impact of an early offer to buy fertilizer at full price is therefore almost as large as the impact of getting the fertilizer at a 50 percent reduced price.

Another feature of the experiment allows the researchers to distinguish whether buying fertilizer when offered at the time of harvest is just another instance of the farmer succumbing to temptation (or because chasing away someone who asks you to buy fertilizer is not pleasant, so if you have the money you may as well just buy it), rather than a rational decision to set aside money for fertilizer use. Before harvest, a field officer visited all farmers to find out when their harvest was. For a group of farmers, they also asked them whether they would be interested in such a product (where ICS sells and delivers fertilizer to them) and if yes, when they should come. This visit takes place in the "hungry season," when farmers have no money. If they just wanted to be nice to the interviewer, but never intended to buy fertilizer (and they knew they would buy it if he were to "tempt" them with it when he arrived at harvest time and they had money), they could tell the interviewer to go away and to come back at planting time. In fact, almost all of the farmers did ask him to come back at some point. However, 44 percent of them asked the field officer to come back at harvest time (rather than at planting time), and most of these farmers eventually bought fertilizer (almost none of the people who asked them to come at planting or top dressing time ended up buying fertilizer). The resulting take-up of the SAFI offer under the "choice of timing" condition was exactly the same as that under the “no choice” condition, suggesting that the decision was a rational one, rather than due to farmers succumbing to pressure at harvest time.

Up to this point, the results seem to vindicate the hypothesis that agents are sophisticated hyperbolic discounters who understand the value of commitment. Interestingly, however, most farmers requested a very rapid delivery of the fertilizer, and ended up storing it for themselves, rather than letting the NGO store it and deliver it when they needed it. Preliminary evidence on adoption suggests that a large fraction of farmers who purchased fertilizer under the SAFI program did use it, suggesting that fertilizer is sufficiently illiquid such that once farmers have it, they manage to hold on to it until they use it. Commitment devices help farmers to save and invest, and they are sufficiently aware of this to take up these devices when offered them. But if farmers can just buy fertilizer and hold on to it, and if they know that they might not end up buying it if they don’t do it early in the season, why don’t they just buy fertilizer on their own immediately after harvest? This seems at odds with the fact that they themselves seem to know that they will not buy fertilizer if it is delivered to them at planting time (since, under the choice condition, almost half of the farmers request a visit at harvest time). Or why doesn’t someone (say, a fertilizer seller) decide to woo customers at harvest time, when they are much more likely to buy it? These questions have prompted a new set of experiments, currently ongoing. In one program, farmers are just reminded at harvest time that if they do not purchase fertilizer right away, they probably will never do it. In one program, they are offered a small discount on fertilizer, with a short deadline (valid only during the immediate post-harvest season). The results of these experiments will help disentangle these possibilities.

II. Lessons

With these three examples, I have tried to show that field experiments have led to substantive learning on key questions of interest to economists. Beyond this, the experience gained in these and other projects has led to methodological insights.

1) The field as a lab

The fertilizer experiments described above were designed to test well-defined theories (production function, learning, and time-inconsistent preferences). As each set of new results came in, there was a constant back and forth between the questions that should be asked and the answers that emerged, but each new program was guided by theory. These experiments are examples of using the design of programs to answer very specific questions guided by theory. The field is used as a lab where variation necessary to test specific ideas is generated experimentally.

Karlan and Zinman (2005) and Bertrand, Karlan, Mullainathan, Shafir, and Zinman (2005) are two related projects which are excellent examples of using the field as a lab. Both projects were conducted in collaboration with a South African lender, giving small loans to high-risk borrowers, with high interest rates. In both cases, the main manipulation started by sending different direct mail solicitation to different people. Karlan and Zinman (2005) set out to test the relative weight of ex-post repayment burden (including moral hazard) and ex-ante adverse selection in lending. In their setup, potential borrowers with the same observable risk were randomly offered a high or a low interest rate in an initial letter. Individuals then decided whether to borrow at the solicitation's "offer" rate. Of those that responded to the high rate, half were randomly given a new lower "contract" interest rate when they actually applied for the loan, while the remaining half continued to receive the rate at which they were offered the loan. Individuals did not know beforehand that the contract rate might differ from the offer rate. The researchers then compared repayment performance of the loans in all three groups. This design allows the researchers to separately identify adverse selection effects and ex-post repayment burden effects (which could be due to moral hazard or sheer financial distress ex post). Adverse selection effects are identified by considering only the sample that eventually received the low contract rate, and comparing the repayment performance of those who responded to the high offer interest rate with those who responded to the low offer interest rate. Ex-post repayment burden effects are identified by considering only those who responded to the high offer rates, and comparing those who ended up with the low offer to those who ended up with the high offer. The study found that men and women behave differently: while women exhibited adverse selection, men exhibited moral hazard. This experiment constitutes a significant methodological advance because it shows how simple predictions from theory can be rigorously tested. This is a very powerful design that allows us to quantify the importance of mechanisms that are at the heart of our understanding of credit markets.

Bertrand et al. (2005) apply the same principle (and the same setup) to a broader set of hypotheses, most of them coming directly from psychology. The experiment is overlaid on the Karlan and Zinman (2005) basic experiment: the offer letters are made to vary along other dimensions, which should not matter economically, but have been hypothesized by psychologists to matter for decision-making, and have been shown to have large effects in laboratory settings. For example, the lender varied the description of the offer, either showing the monthly payment for one typical loan or for a variety of loan terms and sizes. Other randomizations include whether and how the offered interest rate is compared to a "market" benchmark, the expiration date of the offer, whether the offer is combined with a promotional giveaway, race and gender features introduced via the inclusion of a photo in the corner of the letter, and whether the offer letter mentions suggested uses for the loan. The analysis then compares the effect of all these manipulations. While not all of them make a difference, many do, and some of the effects are large and surprising: for example, for male customers, having a photo of a woman on top of the offer letter increased take-up as much as a 1 percent reduction in the monthly interest rate. In some sense, the juxtaposition of the two experiments may be the most surprising: on the one hand, individuals react as "homo economicus" to information—they are sensitive to interest rates and bad risk accepts highest interest rates (at least among women). On the other hand, these effects are present in the same setting where seemingly anodyne manipulations make a large difference.

The two experiments and many others that have been described in this paper illustrate how development economists have gone much beyond program evaluations to use randomization as a tool, and use the field as a "lab." Compared to retrospective evaluations (even perfectly identified ones), field experiments, when the collaboration with the partner is very close, offer much more flexibility and make it possible to give primacy to the hypothesis to test, rather than to the program that happens to have been implemented. With retrospective evaluations, theory is used instrumentally as a way to provide a structure justifying the identifying assumptions (this is more or less explicit depending on the empirical tradition the researchers belong to). With prospective evaluations, it is the experimental design that is instrumental. This gives more power both to test the theory and to challenge it.

The set of fertilizer experiments also illustrates how experiments can be used sequentially, with each set of results providing the inputs for designing a new round of experiments. Such a set of sequential experiments is conducted on the sample population, and in part, on the same panel of farmers, interviewed many times over the course of the experiment. In the process, the researcher successively builds a panel data set on farmers spanning many time units. Though this design remains fairly rare, it offers interesting possibilities in that the experiments become more relevant as the underlying theory becomes more pertinent, and the richness of the data collected allows the researcher to use the data in many other ways than conducting a simple test of the theory. An open question is whether the population becomes affected by staying too long in a panel and being subject to several experiments. This may eventually reduce the external validity of these findings. I am not aware of systematic research on this issue, however.

Lab experiments share this flexibility. Where field experiments are different from lab experiments, however, is that they take place in a context where people are making important decisions, with high stakes. Economists are often suspicious of lab experiments, because it is not clear that behavior observed in the lab would still apply when people make "real" decisions, or whether they would persist over time. Despite their interest, this criticism often applies to some "field experiments" conducted in the U.S., because they take place in very specific, unusual contexts or on rather marginal decisions. Field experiments in development (and some in developed countries as well) have not suffered from this criticism. In the case of the Bertrand et al. (2005) study, for example, the loan sizes average one third of the borrower's income. In the case of Duflo, Kremer, and Robinson (2005), farmers are making agricultural decisions, about which they have plenty of experience. Thus, the lessons from field experiments in development economics are directly relevant to important issues, and also much more likely to generalize.

Working in the field also allows the researchers to calibrate effects against each other. Duflo, Kremer, and Robinson (2005) and Bertrand et al. (2005) quantify the importance of the "non-standard" hypotheses against price effect. They both show that these effects are large. Duflo, Kremer, and Robinson (2005) find that the timing of the offer has almost as large an effect on take-up as that of a 50 percent subsidy. Bertrand et al. (2005) found that any one of the psychological manipulations that had an effect had an effect on take-up of the loan roughly equivalent to that of a 1 percent reduction in the monthly interest rate.

Of course, the fact that the research is conducted with real people facing real decisions also puts limits on what can be done: ethical considerations play an important role, as does the imperative to not propose programs to people that do not make any sense to them. In this sense, the field has fewer options than the lab. However, the external validity of asking people to make decisions that would not make any sense in reality is in any case limited.

2) The relationship between theory and experiments

Field experiments, like lab experiments, are often criticized for lacking external validity (see Basu, 2005): they may be giving the right answer about the behavioral response to an intervention in a particular population, but this is not sufficient to infer that the response would be the same if the intervention was somewhat different, or if the population was somewhat different. The latter is because the experiments are often quite localized and specific in focus. There is no way to generalize the results without recourse to some theory that is external to the experiment (for example, a sense that the treatment effect would be the same for people with similar observed characteristics). As Banerjee (2005) points out, this is clearly correct, but this does not imply that nothing can be learned from a well-executed, empirical exercise: in most of what we do as social scientists, we assume that there is some constancy in nature, so that we can parametrize the contexts according to a limited number of dimensions. This makes it possible (and desirable) to replicate experiments in different contexts. It is always a theory (more or less explicit or well articulated) that will guide the dimensions according to which the experiments will need to be replicated.

To solve both problems of external validity (the specificity in the population and the specificity of the program), several authors, including Mookherjee (2005), have proposed combining experiments with structural models. Attanasio, Meghir, and Santiago (2002) implemented such a strategy on the PROGRESA program. In some sense, the field experiments I have described in this paper subscribe to this approach more radically, since they are generally motivated by the desire to answer one specific well-defined question, and they design the experiment with that in mind. The issue that Attanasio, Meghir, and Santiago (2002) are grappling with in the case of the PROGRESA program is that they are trying to obtain several parameters of interest out of one single experiment which was really a package of interventions (a conditional cash transfer delivered only to women): they then need the theory to provide them with the additional identifying restrictions ex post. The experiments we describe here were set up not to maximize the effect of an intervention, but with a view to understanding the effect that one isolated manipulation would have. In many cases, a stratified design was used in order to identify more than one such relationship. Several "treatments" and different "treatment intensities" can be combined (Karlan and Zinman (2005), for example, combine different interest rates, both ex post and ex ante) to answer more than one question, calibrate treatments against each other, understand their interactions, and get a sense of the "dose response" function. Yet, even as they can be used to test the conceptual foundations of policies, field experiments do not, in general, evaluate a "package" of policies that may be optimal from a policy design point of view. For example, the particular combination of the conditionality of the cash transfer, the way it varies with the age of the child, the fact that the transfers are given to women, and the sheer size of the transfers have presumably been chosen by the promoters of PROGRESA to optimize its effectiveness. Field experiments can be used to test theories, while "traditional" randomized program evaluations can be used to test the effectiveness of more complex policies, combining a variety of policy levers, which have not necessarily been tested, or even implemented, together in the field. Ideally, the results of field experiments and the theories that underlie them would also inform the design of such "combination" policies, so that the two approaches are both policy relevant and complementary.

In field experiments, the structure is imposed ex ante, by the choice of which variations are tested and which are not. However, as we have seen, the results of many of these experiments have challenged the theories they started from and set out to test. There are fundamental reasons why experiments are more likely to generate surprising results than retrospective work. First, as noted previously, randomization frees the experimenter from the need to use theory to justify identification assumptions. Secondly, while it is always possible to reject experimental results on the grounds that the experiment was poorly designed, or failed, when an experiment is correctly implemented (which is relatively easy to ascertain), there is no doubt that it gives us the effect of the manipulation that was implemented, at least in this particular case. It is therefore more difficult to ignore the results even when they are unexpected. An investigator can of course always choose to file the results in a drawer and never mention them again to anyone. It is critical that institutions are developed to avoid this. The FDA requires reporting results of any funded medical trial. Institutions of this type need to be developed for field experiments. Field experiments always start with a proposal which describes the design and the results that are expected. It could (and probably should) be made compulsory for researchers to post their design ex ante and the results corresponding to the design they have posted.

In contrast, non-experimental research designs often leave more room for interpretation and choices, and knowing this, if the initial results accord less well with intuition, a rational Bayesian investigator will give them less weight than she would if they came from an experiment. She will also know that others will be unlikely to believe her study. If she is (like we all are) affected by self-serving biases, she is likely to decide that the initial design was flawed, and may choose to change the specifications, the choice of control variables, etc., until the results accord better with her initial prior.[5] This can happen without any manipulation, just by applying the simple rule of stopping the research when the results "make sense." Consider, for example, the case of learning in the deworming drug cases: if researchers had run the usual specifications and found evidence that children are more likely to take the drug when more of the children in their school take the drug as well, they would have been very likely to just accept this result and publish one more paper confirming positive social learning in this new context. Nobody would have been surprised.

3) What theoretical framework?

[5] Mullainathan (2005) makes this point to highlight that there will be a tendency for evaluation of development programs to find that programs "worked." I think that more generally, researchers will be tempted to find what they want to find. Given the publication bias in the profession (significant results are more interesting to publish), this may be equivalent in most cases.

Field experiments need theory, not only to derive specific testable implications, but to give a general direction of what the interesting questions are. A body of theoretical work allows different results to resonate with each other. Empirical development economists (not only those who conduct field experiments) have greatly benefited from a body of theory that was developed in the 1980s and 1990s, which has been called elsewhere the "poor but neo-classical paradigm." The "poor but neo-classical" paradigm (starting with the work of Stiglitz) incorporates the insights of the economics of information into development. With imperfect information, moral hazard, limited liability, or adverse selection, poverty radically changes the options that an individual has access to: how much someone can borrow depends on his asset position; when insurance options are limited, poor people may be less willing to take any risks. This means that poverty leads to inefficient outcomes, even if everybody is perfectly rational.

This theoretical framework gave a coherent meaning to the empirical results that had been accumulating at the doorstep of the "poor but efficient" framework (Schultz, 1964) and that had been resisted for some time as being inconsistent with theory (for example, the famous farm-size productivity relationship). The initial, theoretical advances opened a new empirical agenda to mainstream economists: the stake was not to accept or reject the hypothesis of "poor but efficient," and with it all the postulates of neo-classical economics; the task of empirical economics shifted to providing evidence for market inefficiencies and the impact of economic policies to alleviate them.

The paradigm "poor but neo-classical" helped define an empirical agenda and structure a vision of the world, even though it often remained implicit in empirical work. It still provides us with a wealth of empirical predictions which could be explicitly tested in the field. Karlan and Zinman (2005), which I discussed above, is a great example of the shape this work can take, but there is little experimental work designed to test these central ideas. The questions are plentiful: How important are dynamic incentives and group lending to the repayment performance in microcredit organizations? If people had access to health or weather insurance, would they undertake riskier, but more profitable, investments? What is the marginal rate of return to capital for small entrepreneurs? Is there evidence of increasing returns to capital over some range? Would increasing the flow of information about prospective borrowers increase lending? Would increased bargaining power of sharecroppers increase agricultural productivity?

One direction in which the work of field experiments needs to go in the future is thus to exploit more fully this powerful theoretical framework to come up with hypotheses to test in the field.

However, we also need to deal with the fact that the results of many of these experiments have challenged this framework. In the absence of a well-funded alternative frame of analysis, these rejections appear now as a collection of random results that do not fit very well within any existing theory, and that we don't necessarily fully understand. This makes it difficult to generalize results and give them meaning, as some of the critics of randomized evaluation have pointed out. However, criticizing the experiments on this ground, like many have done, is a little bit like shooting the messenger. One may instead want to accept the message that they deliver: that we need to work on a new theoretical framework that can accommodate these results and predict new ones.

Banerjee (2005) identifies this as the "new challenge to theory." According to him (a central contributor to the "poor but neo-classical" framework), the challenge is to form a new body of theory which can be used as a general framework to make sense of the disparate results emerging from the field. For this, he argues, "we need to give up trying to defend the existing theory (which has been incredibly successful in many ways) against the onslaught of seemingly random results that are coming out of the field experiments." I see these results less as a challenge than an opportunity. It was not the remit of the "poor but neo-classical" framework to explain the entire world, and in this sense, it does not need to be "defended" for not being able to explain everything.

While the empirical work continues to explore the relevance and the limits of the "poor but neo-classical" framework, a direction in which the theoretical work needs to be going, therefore, is to start working on a theoretical framework that can accommodate the new results; this is exactly what theorists did when the "poor but neo-classical" framework incorporated and replaced the "poor but rational" one. This theory is lacking at the moment. While many experiments use insights gleaned from behavioral economics or psychology to design tests and interventions, the work of organizing these insights into a coherent framework that applies to development has not taken place. Behavioral economics, in particular, has not yet produced a coherent unifying theory. The "theories" to be tested sometimes look more like a collection of anomalies. Moreover, faithfully applying the theories developed for developed countries to the analysis of the decisions of the poor in developing countries would, however, be making the same mistake as the "poor but efficient" proponents and failing to recognize the central insight of the "poor but neo-classical" line of research. Trying to reduce the behavior of a Kenyan farmer who does not use fertilizer and that of an American employee who does not contribute to his 401k to the same model may be as fruitless as trying to convince oneself that Guatemalan farmers are on the efficiency frontier. The same limitation in cognitive ability or self-control that affects the rich may also affect the poor, and has different implications for them. As Mullainathan (2005) points out, the point is not to say that the poor are just particularly irrational, but to recognize that the same failure of rationality may have dramatically different consequences for the poor. Being poor almost certainly affects the *way* people think and decide. Perhaps when choices involve the subsistence of one's family, trade-offs are distorted in different ways than when the question is how much money one will enjoy at retirement. Pressure by extended family members or neighbors is also stronger when they are at risk of starvation.

What is needed is a theory of how poverty influences decision-making, not only by affecting the constraints, but by changing the decision-making process itself. That theory can then guide a new round of empirical research, both observational and experimental.

III. Conclusion: open questions

I'll conclude by briefly outlining two open areas where research (experiments and theory) are particularly needed: these are areas that are both of tremendous practical importance and of great interest to research.

The first can broadly be named the question of behavior change: why are people not doing things that are obviously good for them or their children (even when they love their children), such as using a condom to protect themselves from HIV/AIDS, taking their TB medicine, immunizing their children, getting free antenatal checkups, using an oven with a chimney to avoid filling up the room with smoke when cooking, etc. This is clearly a phenomenon that is common to developed and developing countries, but the consequences are often so dire in developing countries that this is, to paraphrase Lucas, a problem that it is almost impossible to let go of when one starts thinking about it. Behavioral economists are studying similar problems in developed countries, with particular attention to the question of savings. Their approach has been in some cases akin to the approach we described above. For example, Thaler and Benartzi (2003) developed a financial product, "Save More Tomorrow," which allowed a new employee to save a fraction of future increments in their salaries in their 401k. This is precisely a product that would appeal to a hyperbolic discounter. Though it was not evaluated formally (there was no experiment), the program appeared to be extremely successful, and has now been adopted by many companies. Many NGOs in developing countries are engaged in trying to solve exactly these problems. Collaborating with them to evaluate their approach or design and evaluate new approaches could help build a body of effective interventions. By being open-minded about what will work (beyond information and incentives), one can make progress both in understanding behavior and in improving lives considerably.

The second area does not involve solving intra-person problems, but interpersonal ones. Development economists have always stressed the importance of institutions, and recently, the study of institutions has re-emerged as one of the central questions in development (see Pande and Udry (2005) in this volume). A central reason for underdevelopment is the lack of institutions that favor cooperation and social behavior. The central practical question then becomes: what to do about poor institutions? Should we just write off those countries which are plagued with (often historically inherited) poor institutions, or should we instead work on ways to get things done in these environments (with a view, perhaps, to arrive at institution change eventually). Most countries with very poor institutions function to some extent. Absenteeism rates are extremely high in schools and health services, but what may be surprising is that nurses or teachers actually come at all in the absence of any sort of punishment for delinquent behavior (Chaudhury et al. (2005)). Understanding how to harness people's intrinsic motivation and social preferences may help improve the day-to-day functioning of countries where institutions are in disarray. There is nothing in this that is particularly new: as Ray (1998) points out in the introduction to his textbook, development economics is in large part the study of indigenous, informal institutions that emerge to palliate the absence of well-functioning formal institutions. This may just demand researchers to be a bit more imaginative in thinking about what can motivate people.

In all these cases, the economist goes beyond a purely positive role and does not shy from assuming a normative position. This was already advocated by Banerjee (2002). Working in developing countries makes one acutely aware of how much "slack" there is in the world, and how small interventions can make large differences. But if economists are normative, it becomes critical that they rigorously evaluate their propositions, since, like most people feeling around for the light switch, they are likely to make mistakes. This makes the experimental approach indispensable, both as a practical and as a scientific tool.

REFERENCES

Ainslie G. (1992). *Picoeconomics*. Cambridge: Cambridge University Press

Ashraf, Nava, Dean S. Karlan, and Wesley Yin (2006). "Tying Odysseus to the Mast: Evidence from a Commitment Savings Product in the Philippines," *Quarterly Journal of Economics*, forthcoming

Angrist, Josh and Victor Lavy (2002). "The Effect of High School Matriculation Awards: Evidence from Randomized Trials," NBER Working Paper No. 9389

Attanasio, Orazio, Costas Meghir, and Ana Santiago (2002). "Education Choices in Mexico: Using a Structural Model and a Randomized Experiment to Evaluate PROGRESA," MIMEO, Inter-American Development Bank

Banerjee, Abhijit (2002). "The Uses of Economic Theory: Against a Purely Positive Interpretation of Theoretical Results," MIMEO, MIT

Banerjee, Abhijit (2005). "'New Development Economics' and the Challenge to Theory," *Economic and Political Weekly*, October 1, 2005

Banerjee, Abhijit and Ruimin He (2003). "The World Bank of the Future," *American Economic Review, Papers and Proceedings*, 93(2): 39-44

Banerjee, Abhijit, Angus Deaton, and Esther Duflo (2004). "Health Care Delivery in Rural Rajasthan," *Economic and Political Weekly*, 39(9), 944-949

Banerjee, Abhijit and Esther Duflo (2005). "Addressing Absence," *Journal of Economic Perspectives*, forthcoming

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

Banerjee, Abhijit and Sendhil Mullainathan (2005). "Motivation and Poverty," MIMEO, MIT

Basu, Kaushik (2005). "The New Empirical Development Economics: Remarks on Its Philosophical Foundations," *Ecomic and Political Weekly*, October 1, 2005

Bertrand, Marianne, Dean Karlan, Sendhil Mullainathan, Eldar Shafir, and Jonathan Zinman (2005). "What's Psychology Worth? A Field Experiment in the Consumer Credit Market," Yale University Economic Growth Center Discussion Paper No. 918

Besley, Timothy and Anne Case (1994). "Diffusion as a Learning Process: Evidence from HYV Cotton," RPDS, Princeton University, Discussion Paper No. 174

Binswanger, Hans (1980). "Risk Attitudes of Rural Households in Semi-Arid Tropical India," *American Journal of Agricultural Economics 62*: 395-407

Chaudhury, Nazmul, Jeffrey Hammer, Michael Kremer, Karthik Muralidharan, and F. Halsey Rogers (2005). "Teacher Absence in India: A Snapshot," *Journal of the European Economic Association* 3(2-3), April-May

Conley, Timothy and Christopher Udry (2005). "Learning About a New Technology: Pineapple in Ghana," MIMEO, Yale

Duflo, Esther (2004). "Scaling Up and Evaluation" in *Accelerating Development*, edited by Francois Bourguignon and Boris Pleskovic. Oxford, UK and Washington, DC: Oxford University Press and World Bank

Duflo, Esther, Rachel Glennerster, and Michael Kremer (2005). "Randomization as a Tool for Development Economists," MIMEO, MIT

Duflo, Esther and Rema Hanna (2005). "Monitoring Works: Getting Teachers to Come to School," NBER Working Paper No. 11880, December

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), edited by Osvaldo Feinstein, Gregory K. Ingram, and George K. Pitman. New Brunswick, NJ: Transaction Publishers, pp. 205-232

Duflo, Esther, Michael Kremer, and Jonathan Robinson (2005). "Understanding Fertilizer Adoption: Evidence from Field Experiments," Mimeo, MIT

Duflo, Esther and Emmanuel Saez (2002). "Participation and Investment Decisions in a Retirement Plan: The Influence of Colleagues' Choices," *Journal of Public Economics*, 85(1): 121-148

Duflo, Esther and Emmanuel Saez (2003). "The Role of Information and Social Interactions in Retirement Plan Decisions: Evidence from a Randomized Experiment," *Quarterly Journal of Economics*, 118(3): 815-842

Foster, Andrew and Mark Rosenzweig (1995). "Learning by Doing and Learning from Others: Human Capital and Technical Change in Agriculture," *Journal of Political Economy*, 103: 1176-1209

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

Holmstrom, Bengt and Paul Milgrom (1991). "Multitask Principal-Agent Analysis: Incentive Contracts, Asset Ownership and Job Design," *Journal of Law, Economics and Organization*, 7: 24-52

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

Karlan, Dean and Jonathan Zinman (2005). "Observing Unobservables: Identifying Information Asymmetries with a Consumer Credit Field Experiment," MIMEO, Yale

Kling, Jeffrey and Jeffrey Liebman (2005). "Experimental Analysis of Neighborhood Effects on Youth," NBER Working Papers No. 11577

Kremer, Michael and Daniel Chen (2001). "An Interim Report on a Teacher Attendance Incentive Program in Kenya," MIMEO, Harvard

Kremer, Michael and Edward Miguel (2003). "Networks, Social Learning, and Technology Adoption: The Case of Deworming Drugs in Kenya," MIMEO, Harvard

Kremer, Michael, Edward Miguel, and Rebecca Thornton (2004). "Incentives to Learn," NBER Working Paper No. 10971

Laibson D. (1997). "Golden Eggs and Hyperbolic Discounting," *Quarterly Journal of Economics*, 62: 443-478

Manski, Charles (1993). "Identification of Exogenous Social Effects: The Reflection Problem," *Review of Economic Studies*, 60: 31-542

Mookherjee, Dilip (2005). "Is There Too Little Theory in Development Economics?" *Economic and Political Weekly*, October 1, 2005

Mullainathan, Sendhil (2005). "Development Economics Through the Lens of Psychology" in

Annual World Bank Conference in Development Economics 2005: Lessons of Experience, edited by Francois Bourguignon and Boris Pleskovic. Oxford, UK and Washington, DC: Oxford University Press and World Bank

Munshi, Kaivan (2004). "Social Learning in a Heterogeneous Population: Technology Diffusion in the Indian Green Revolution," *Journal of Development Economics*, 73(1): 185-215

Munshi, Kaivan and Jacques Myaux (2002). "Social Norms and the Fertility Transition," *Journal of Development Economics*, forthcoming

Olken, Benjamin (2005). "Monitoring Corruption: Evidence from a Field Experiment in Indonesia," NBER Working Paper No. 11753

Pande, Rohini and Christopher Udry (2005). "Institutions and Development: A View from Below," in *Advances in Economics and Econometrics: Ninth World Congres*, edited by Richard Blundell, Whitney Newey, and Torsten Persson. Cambridge, UK: Cambridge University Press

Ray, Debraj (1998). *Development Economics*. Princeton, NJ: Princeton University Press

Schultz, Theodore W. (1964). *Transforming Traditional Agriculture*. New Haven, CT: Yale University Press

Strotz, R. (1956). "Myopia and Inconsistency in Dynamic Utility Maximization," *Review of Economic Studies*, 23:165-180

Thaler, Richard and Shlomo Benartzi (2004). "Save More Tomorrow: Using Behavioral Economics to Increase Employee Saving," *Journal of Political Economy*, 112(1): 164-187

Thornton, Rebecca (2005). "The Demand for and Impact of Learning HIV Status: Evidence from a Field Experiment," MIMEO, Harvard

Todd, Petra and Kenneth I. Wolpin (2004). "Using a Social Experiment to Validate a Dynamic Behavioral Model of Child Schooling and Fertility: Assessing the Impact of a School Subsidy Program in Mexico," WP 03-022, University of Pennsylvania

Townsend, Robert (1994). "Risk and Insurance in Village India," *Econometrica*, 62(4):539-591

Udry, Christopher (1994). "Risk and Insurance in a Rural Credit Market: An Empirical Investigation in Northern Nigeria," *Review of Economic Studies*, 61(3): 495-526