

# The Experimental Approach to Development Economics

Abhijit V. Banerjee and Esther Duflo

Department of Economics and Abdul Latif Jameel Poverty Action Lab, Massachusetts Institute of Technology, Cambridge, Massachusetts 02142; email: banerjee@mit.edu, eduflo@mit.edu

Annu. Rev. Econ. 2009.1:151-178. Downloaded from www.annualreviews.org by 24.37.227.27 on 12/19/10. For personal use only.

Annu. Rev. Econ. 2009. 1:151–78

First published online as a Review in Advance on April 21, 2009

The *Annual Review of Economics* is online at [econ.annualreviews.org](http://econ.annualreviews.org)

This article's doi:  
10.1146/annurev.economics.050708.143235

Copyright © 2009 by Annual Reviews.  
All rights reserved

1941-1383/09/0904-0151\$20.00

## Key Words

randomized experiments, development economics, program evaluation

## 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, owing to the close collaboration between researchers and implementers, they allow the estimation of parameters that would not otherwise be possible to evaluate. We discuss the concerns that have been raised regarding experiments and generally conclude that, although real, they are often not specific to experiments. We conclude by discussing the relationship between theory and experiments.

## 1. INTRODUCTION

The past few years have seen a veritable explosion of randomized experiments in development economics. At the fall 2008 NEUDC conference, a large conference in development economics attended mainly by young researchers and PhD students, 24 papers reported on randomized field experiments, out of the 112 papers that used microeconomics data (laboratory experiments excluded). This is up from 4 in 2004. At the fall 2008 BREAD conference, the premier conference on development economics, 4 of the 8 papers invited were randomized field experiments. Three out of the 6 papers using microeconomic data from developing countries published in 2008 or forthcoming in the *Quarterly Journal of Economics* involve randomized assignment. The enthusiasm is not limited to academia. At the World Bank, there are 67 ongoing randomized evaluations (out of 89 ongoing program evaluations) in the African region alone.

Perhaps inevitably, this progress has also generated 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 written by James Heckman more than 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 define the strengths and limitations of randomized experiments, and in clarifying where the field needs to go next. However, we 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 then return to the various criticisms, in part to clarify and qualify them and 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.

## 2. THE PROMISE OF EXPERIMENTS

Experimental research in development economics, like earlier research in labor economics, health, and education, started from a concern about the reliable identification of program effects in the face of complex and multiple channels of causality. In general, participants to a program differ from nonparticipants in many ways, and we have no information on how they would have fared had they not been participating. This makes it difficult to separate the “causal effect” of the program (i.e., for a given participant, the difference between the outcome he experienced under the program and the outcome he would have experienced if he had not been participating) from other factors. A central problem is selection, the fact that participants may be systematically different from nonparticipants. Although the treatment effect for each person cannot be identified, experiments make it possible to vary one factor at a time and therefore provide internally valid estimates of the average treatment effect for a population of interest. [For detailed discussions of the evaluation problem, see Heckman & Vytlačil (2008a) and Imbens & Woolridge (2008).]

The experimental work in the mid-1990s (e.g., Glewwe et al. 2004, 2009; Banerjee et al. 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?

The motivating theoretical framework was thus very simple, but this research produced a number of surprising results, both negative and positive: For example, improving access to textbooks from one per four or more students to one per every two does not affect average test scores (Glewwe et al. 2009); halving the teacher-student ratio also had no effect (Banerjee et al. 2005). However, a study of treatment for intestinal worms in schools in Kenya (Miguel & Kremer 2004) showed that a deworming treatment that costs 49 cents per child per year can reduce absenteeism by 25%. In part, this is because of externalities: Worms are transmitted when children walk barefoot in places where other children who are infected by worms have defecated. As a result, in terms of increasing attendance, deworming is nearly 20 times as effective as hiring an extra teacher (the cost for an extra child-year of education was \$3.25 with deworming, whereas the cost was approximately \$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 be very helpful—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 (2005) compares the cost per extra child-year of education induced across an array of different strategies. The costs vary widely, between \$3.50 per extra child-year for deworming and \$6000 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 ranges from \$3.25 to more than \$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; implementing organizations were not much better informed. For example, the nongovernmental organization (NGO) that financed the deworming intervention was also initially enthusiastic about giving children school uniforms, even though a randomized evaluation showed that the cost of giving children a free uniform worked out to be \$100 per extra child-year of schooling.

Several important conclusions emerged from this experience. First, effective policymaking requires making judgments about the efficacy of individual program components—without much guidance from a priori knowledge. Second, however, it is also difficult to learn about these individual components from observational (i.e., nonexperimental) 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 good reasons, involves making multiple changes at the same time. Although there are exceptions, 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 provides 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 (to make use of the results), it is worth doing multiple experiments at the same time on the same population to 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 grade-1 class sizes in Kenya. When researchers worked with the school system to set up an evaluation of the program, they did not just assign the entire program to the randomly selected treatment schools (Duflo et al. 2008a). Instead, they introduced two additional dimensions of variation: (a) training of the school committee that received the money to monitor the extra teacher and (b) tracking by prior achievement. Using this design, researchers can estimate the impact of class-size reduction without changing pedagogy; the relative merit of young, extra teachers on short contracts versus 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 et al. (2005), albeit in a different context, the study did 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 et al. (2007) (remedial education and computer-assisted learning), Duflo et al. (2006) and Dupas (2007) (various HIV-AIDS prevention strategies among adolescents), Banerjee et al. (2009) (information and mobilization experiments in primary schools in India), Banerjee et al. (2008) (demand and supply factors in improving immunization rates in India), and Gine et al. (2008) (two strategies to help smokers quit smoking).

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 because experimental results are often surprising and therefore require further clarification. Duflo et al. (2008c,d) reflects exactly such an iterative process, where a succession of experiments on fertilizer use was run over a period of several years. Each result prompted the need to try out a series of new variations to better understand the results of the previous one.

In addition, 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 to understand what part of the broad program works.<sup>1</sup> Policy experiments often stop at the first step. One example is the popular PROGRESA-Oportunidades program in Mexico, which combined a cash transfer to women in poor families that was conditional on “good behavior” (e.g., investments in education and preventive health) and some upgrading of education and health facilities. The program has been replicated in many countries, often along with a randomized evaluation (Fizbein & Schady 2009). However, only an ongoing study in Morocco formed and compared different treatment groups so that researchers could 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 (e.g., attendance requirements are verified only by teachers), and two groups receive

---

<sup>1</sup>The opposite approach, i.e., going from one intervention at a time to the full package, makes also sense when your priors are that some combination will work, whereas the alternate is better when you are generally skeptical.

stricter variants of the conditionality (in one group, children's attendance is supervised by inspectors; in the other, it is verified daily with a fingerprint recognition device).

Although all this seems obvious in retrospect, it was only after the first few experiments that both researchers and the implementing organizations fully appreciated the significance of such a design. 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 ongoing 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 United States or Canada where, with a few important exceptions (e.g., Angrist et al. 2009), the programs to be evaluated are mainly chosen by the implementing agencies and the researchers are evaluators only.

From the point of view of the researchers, this design offered the possibility of moving from the role of the evaluator to the role of a coexperimenter, which included 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, thus drawing upon his knowledge of what else was known and the received theory. For example, when Seva Mandir, an NGO in Rajasthan, India, with whom we have had a long-standing relationship, was interested in improving the quality of their informal schools, their initial idea was to implement a teacher incentive program based on test scores. However, they were persuaded by the results from Glewwe et al. (2003) that showed that teacher incentives could result in teaching to the test or other short-run manipulations of test scores. They then decided to implement an incentive program based on teacher presence. To measure attendance in very sparsely populated areas where schools are difficult to access, Duflo and Hanna (Duflo et al. 2007) proposed the use of cameras with date and time stamps. Although Seva Mandir was initially 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 were striking (Duflo et al. 2007): Teacher absence dropped from 40% to 20% while students' performance 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 of their schools immediately, they decided to continue it in the schools where it 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 send every child a diary to write in every day based on work done in school. On days when the student or the teacher was absent, the diary was to remain blank or the date was to be crossed out. Parents were supposed to look at the diary every week. The hope was that they would register the extent of teacher and child absence. This approach, it turned out, did not succeed: Parents started with such a low opinion of school that the diary tended to persuade them that something was happening, regardless of number of absences. Indeed, parents of diary schools had a higher opinion than did those of nondiary schools, and there was no impact on teacher presence. However, the diaries were popular with both students and teachers, and their use induced teachers to work harder. 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, because this was not a hypothesis put forward in the initial experimental design, it may just be a statistical accident. Thus, Seva Mandir will now put cameras in all schools (after several years, they continue to have a large impact on presence and tests scores), while they also 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.<sup>2</sup> 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 et al. (2008) and Angrist & Lavy (2009)] or parents (as in Kremer et al. 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 designed a program in the slums of Delhi where students (or their parents) were provided incentives (in the form of toys or money) based on the child's improvement in reading. He found that, for initially weak students, rewarding the child is more effective than rewarding the parents in terms of improving test scores, whereas the opposite is true for initially strong students. The ability to vary who receives the incentives within the same context and in the same experiment is what made this study possible.

Experiments are thus emerging as a powerful tool for testing theories. Although the theories to be tested are different, the motivation of the recent literature in experimental development economics is similar to that of the first generation of experiments in the United States, which were designed to identify well-designed parameters (i.e., income and substitution effect in the negative income tax experiments, moral hazard in the Rand health insurance experiment, etc.). Interventions are being designed and evaluated not only to show the average treatment effect for a particular policy or program, but also to allow identification of specific economic parameters. One example is a project conducted by Karlan & Zinman (2005) in collaboration with a South African lender that gives small loans to high-risk borrowers at high interest rates. The experiment was designed to test the relative weights of *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 who apply at the higher rate, half are randomly offered a new lower contract interest rate when they are actually given the loan, whereas the remaining half continue at the offer rate. Individuals did not know *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

<sup>2</sup>This flexibility is, of course, not boundless. Ethical concerns (supervised by universities' internal review boards) and the constraint of working with an implementing organization do limit the set of questions you can ask, relative to what one can do in a lab experiment. Not everything can be tested, and not everyone wants to be experimented on. However, the extra realism of the setting is an enormous advantage. It should also be noted that the lower cost of the programs and working with NGO partners greatly expand the feasible set of experiments in development, compared with what has been feasible in the United States.

offer interest rate in the population that received the same low contract rate allows the identification of the adverse selection effect; comparing those who faced the same offer rate but differing contract rates identifies the repayment burden effect.

The study found that women exhibit adverse selection but men exhibit moral hazard. The fact that this difference was unexpected poses something of a problem for the paper (Is it a 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 been replicated in several different studies. Ashraf et al. (2007) and Cohen & 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 the offer price has on keeping the actual price fixed identifies the screening effect, whereas the variation in the actual price (with a fixed offer price) pins down the sunk cost effect. Ashraf et al. (2007) studied this for a water-purification product, whereas Cohen & Dupas (2007) focused on bed nets. Neither study shows 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) that aimed to obtain distinct wage and income variations to estimate income and substitution effects that were not available in observational data (Heckman 1992).

Other examples of this type of work are the experiments designed to assess whether there is a demand for commitment products, which could be demanded by self-aware people with self-control problems. Ashraf et al. (2006) worked with microfinance institutions in the Philippines to offer their clients a savings product that let them choose to commit not to withdraw the money before a specific time or amount goal was reached. Gine et al. (2008) worked with the same organization to invite smokers who wanted to quit to put a “contract” on themselves: Money in a special savings account would be forfeited if they failed a urine test for nicotine after several weeks. Both cases were designed by the economists to solve a real-world problem, but they also 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: Because the cases were experimental in nature, the partners were happy to try them out first 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 et al. 2009, who set up an experiment to understand the structure of corruption in the process of obtaining a driving license in Delhi. They recruited people who are aiming to get a driving license and set up three groups, one that receives a bonus for obtaining a driving license quickly, one that gets free driving lessons, and a control group. They found that those in the “bonus” group get their licenses faster, but those who get the free driving lessons do not. They also found that those in the bonus group are more likely to pay an agent to get the license (who, they conjecture, bribes someone). They also found that the applicants who hired an agent were less likely to have taken a driving test before getting a driving license. Although they did not appear to find that those in the bonus group who get licenses are systematically less likely to know how to drive than those in the control group (which would be the litmus test that corruption does result in an inefficient allocation of

driving licenses), this experiment provides suggestive evidence that corruption in this case does more than “grease the wheels” of the system.

The realization that experiments are a readily available option has also spurred creativity in measurement. In principle, there is no automatic link between careful and innovative collection of microeconomic data and the experimental method. And, indeed, there is a long tradition in development economics to collect data specifically designed to test theories, and both the breadth and the quantity of microeconomic data collected in development economics have exploded in recent decades, not only in the context of experiments.

However, the specificity that experiments have, which is particularly prone to encourage the development of new measurement methods, is high take-up rates and a specific measurement problem. In many experimental studies, a large fraction of those who are intended to be affected by the program are actually affected. This means that the number of units on which data needs to be collected to assess the impact of the program does not have to be very large and that data are typically collected especially for the purpose of the experiment. Elaborate and expensive measurement of outcomes is then easier to obtain than in the context of a large multipurpose household or firm survey. By contrast, observational studies must often rely for identification on variation (policy changes, market-induced variation, natural variation, supply shocks, etc.) that cover large populations, requiring the use of a large data set often not collected for a specific purpose. This makes it more difficult to fine-tune the measurement to the specific question at hand. Moreover, even if it is possible *ex post* to do a sophisticated data collection exercise specifically targeted to the question, it is generally impossible to do it for the preprogram situation. This precludes the use of a difference-in-differences strategy for these types of outcomes, which again limits the incentives.

Olken (2007) is one example of the kind of data that were collected in an experimental setting. The objective was to determine whether audits or community monitoring were effective ways to curb corruption in decentralized construction projects. Getting a reliable measure of actual levels of corruption was thus necessary. Olken focused on roads and had engineers dig holes in the road to measure the material used. He then compared that with the level of material reported to be used. The difference is a measure of how much of the material was stolen, or never purchased but invoiced, and thus an objective measure of corruption. Olken then demonstrated that this measure of “missing inputs” is affected by the threat of audits, but not, except in some circumstances, by encouraging greater attendance at community meetings.

Another example of innovative data collection is found in Beaman et al. (2009). The paper evaluates the impact of mandated political representation of women in village councils on citizens’ attitude toward 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 the 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). Although those tests are frequently used by psychologists, and their use has also been advocated by economists (Bertrand et al. 2005), they had not been implemented in a field setting in a 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 toward women, inspired by political scientists. The respondents listen to a

speech, supposedly given by a village leader, delivered by either a 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 unreserved villages.

These are only two examples of a rich literature. Many field experiments embed small lab experiments (dictator games, choices over lotteries, discount rate experiments, public good games, etc). For example, in their evaluation of the Columbia Conditional Cash Transfer Program, the team from the Institute for Fiscal Studies included public goods, risk sharing, and coalition formation games as part of the data collection (Attanasio et al. 2008a).

### 3. CONCERNS ABOUT EXPERIMENTS

As we mention above, the concerns about experiments are not new. However, many of these concerns 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 overinterpreting 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 each one. Note that, although some of these issues are specific to experiments (we point these out along the way), most of these concerns (external validity, the difference between partial equilibrium and market equilibrium effects, nonidentification of distribution of effect) are common to all microevaluations, both with experimental and nonexperimental methods. They are more frequently brought to the forefront when discussing experiments, which is likely because most of the other usual concerns are taken care of by the randomization.

#### 3.1. Environmental Dependence

Environmental dependence is a core element of generalizability (or external validity). It asks the question, Would we get the same result if we carried out the same experiment in a different setting, or more exactly, would the program that is being evaluated have the same effect if it were 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 environment where the program is evaluated 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 help us here only to some extent—simple economics suggests that uniforms will have an effect only 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.

Second, and more specific to experiments in development economics, which have often been conducted with NGOs, is the issue of implementer effects. That is, the smaller the implementing organization, the greater the concern that the estimated treatment effect reflects the unique characteristics of the implementer. 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 (e.g., How big was the evaluated piece relative to what they do? How was the implementing team selected? What decided the choice of location?). 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 so that its implementation routinely gets delegated to a large number of more or less self-sufficient individual implementing teams.

All of this, however, is very loose and highly subjective (What is large enough? How self-sufficient?). 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. The supplemental teaching (“balsakhi”) program evaluated by Banerjee et al. (2007) was 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 et al. (2006) obtained an impact of a combination of deworming and iron supplementation on school attendance in north India similar to what Miguel & Kremer (2004) found in Kenya. Likewise, Bleakley (2007) found similar results using natural data from the southern United States in the early part of the twentieth century using a natural experiment approach. The PROGRESA-Oportunidades 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). [For a discussion of the original PROGRESA evaluation and subsequent replications, see Fiszbein & Schady (2009).] The results, analyzed by different teams of researchers in different countries, were consistent across countries.

Other results turn out not to be replicable: An information campaign that mobilized parents’ 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 overcrowded had a positive impact on learning outcomes in Kenya but not in India

(Banerjee et al. 2009, Duflo et al. 2008a). A similar intervention that sought to energize Health Unit Management Committees in Uganda also reported a massive impact on hard to affect outcomes such as infant mortality (Bjorkman & Svensson 2007).

In addition to pure replication, cumulative knowledge is generated from related experiments in different contexts. The analytical review by Kremer & Holla (2008) of 16 randomized experiments of price elasticity in health and education is a nice example. We return to these results in more detail below, but the key point here is that these experiments cover a wide range of education and health goods and services in several countries. A 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). Although they are not strictly replications of each other, this shows the value of cumulative knowledge in learning about one phenomenon.

However, more replication research is needed. Some worry that there are little incentives in the system to carry out replication studies (because 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 a recent development, so we do not know how big a problem this may be, but 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 under way. 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 & Zinman (2008) are currently under way in Ghana, Peru (in two separate locations with two different partners), Mexico, and the Philippines (in three separate locations with two different partners). Microcredit impact evaluations are happening simultaneously in Morocco, urban India, the 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, the 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 ongoing ultra-poor experiments, the Ford Foundation is funding all of them, in an explicit attempt to gain a better 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 in 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. The insights of the literature on heterogeneous treatment effects, which we discuss below, can be applied here: First, the different site dummies can be treated as covariates in a pooled regression; nonparametric tests of heterogeneity (e.g., Crump et al. 2009) can be performed. 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 is to test whether the treatment

effect conditional on the covariates is equal for all the site dummies (Heckman et al. 2010). The point is 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. Although identifying a particular quasi-experiment that provides a source of identification for the effect of a particular program may be possible, 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 nonconfoundedness (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 overidentification test), and another explanation 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.<sup>3</sup> In this sense, it is suggested, there is a trade-off between the more “internally” valid randomized studies and the more “externally” valid observational studies, yet, this is 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 that 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 the set of things that happened to have been pooled together when constructing the aggregate data.

There is also a more subtle issue about generalizations that arises even when we evaluate 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 try first 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 to look at how those in the program perform relative to those not in the program. However, once we match like with like, either almost everyone who is in a particular matched group may be a program participant or everyone may be a nonparticipant. There are several methods to deal with this lack of overlap between the distribution of participants and nonparticipants (Heckman et al. 1997a, 1998; Rubin 2006—also see a review in Imbens & Wooldridge 2008), but in all

---

<sup>3</sup>Note that not all randomized experiments are small scale. For example, the mandated representation programs we mention above were implemented nationwide in India. Whereas Duflo & Chattopdhyay (2004) originally looked at only two (very different) states, Topalova & Duflo (2003) extended the analysis to all the major Indian states.

cases, the estimate will be entirely driven by the subgroups in the population where, even after matching, there are both enough participants and nonparticipants, and these subgroups could be entirely nonrepresentative. Even though 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 unobservables. In the words of Imbens & Wooldridge (2008), “a potential feature of all these methods [that improve overlap between participants and nonparticipants] 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 also present in observational studies. By contrast so 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 & 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.

“Randomization bias” is one issue regarding the ability to generalize that is specific to evaluation: The fact that the program is evaluated using a randomized evaluation changes the way that actors behave in the experiment. One form of randomization bias is the Hawthorne effect or the John Henry effect: The behavior of those in the treated or the control group changes because they know that the program is being evaluated, so although the estimate of the effect of the program is internally valid, it has no relevance outside of the experiment (Heckman & Vytlačil 2008b). Hawthorne effects are, in principle, a problem in any setting where participants are studied and are not specific to a given experiment.<sup>4</sup> Social scientists worry about reporting bias (for example, because people want to give a certain impression of themselves to a field officer). However, in an experiment, subjects who know that they are being studied may purposefully try to make the treatment a success or a failure. (We discuss Hawthorne effects in more detail in Section 3.3.)

Another, more subtle form of randomization bias is discussed by Heckman (1992). He points out that that, in the JTPA experiment, not every site agreed to participate, and in particular, some sites specifically refused because there was a randomization. Those sites may be different and have unique treatment effects. Experiments in development economics tend to be conducted with a variety of partners, but it is true that not every NGO or government is willing to participate in a randomized evaluation. If randomized evaluations can be carried out only 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)—i.e., compliance at the level of the organization—and one that is difficult to refute, because 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 of an issue as randomized evaluations gain wider acceptance: Evaluation projects have been completed with international NGOs, local governments, and an array of local NGOs. This situation will continue to improve if randomized evaluation comes to be recommended by most donors, as the willingness to comply with randomization will then be understood not to set organizations apart.

<sup>4</sup>In fact, the original Hawthorne effect happened during “experiments” in workplace conditions that were not randomized.

Such is already happening. In particular, many World Bank researchers and implementers are working with developing-country governments to start an ambitious program of program evaluation.<sup>5</sup> For example, the Africa Impact Evaluation Initiative is supporting (with money and technical capacity building) a number of African governments to start randomized evaluation on various topics in Africa. Currently, 67 randomized evaluations are ongoing within this program, covering five themes: education, malaria, HIV-AIDs, accountability, and transport. An evaluation recently completed under this umbrella, the AGEMAD, a school reform initiative in Madagascar (World Bank 2008), demonstrates the willingness and ability of a ministry of education to implement a randomized evaluation project, when given the necessary support and encouragement by a major donor.

A more serious issue, in our experience, is the related fact that what distinguishes possible partners for randomized evaluations are 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 of 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 from 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 are collected specifically 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, which is politically more sensitive to evaluate than are pilot programs, because the former is usually well publicized, so countries may be strategic with respect to the choice of programs to evaluate. In other cases, regular, large-scale surveys (such as the National Sample Survey in India, the Susenas in Indonesia, etc.) can be used. But not all developing countries have them (though data sets such as the Demographic and Health Surveys, which are available for many countries, have certainly ameliorated the issue). Thus, a potential bias (although distinct from that of randomized evaluation) also exists in the types of countries and programs that can be evaluated with observational data. The point here is not that generalizability is not an issue for the experimental/quasi-experimental approach, but that it is not obviously less of an issue for any other approach.

### 3.2. Compliance Issues

Above, we make the point that a high compliance rate makes it easier to interpret the instrumental variables estimate of the “treatment on the treated” estimates and, therefore, to generalize the results to other environments. 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 location level, the take-up rate is high, often 100%. It should be

---

<sup>5</sup>Francois Bourguignon and Paul Gertler, when they were the chief economist and the chief economist of the human development network at the World Bank, respectively, played key roles in encouraging evaluations.

emphasized that this means only that the treated sample is likely to be a random subset of the set of locations that were selected for the program. The actual individuals who benefited from the treatment are 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 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 a “treatment” village as the instrument is the policy parameter of interest.

Heckman (1992) was specifically concerned with the interpretation of the results of randomized experiments in the United States where individuals were offered the option to participate in a job training program. The fact that take-up was low and potentially highly selected implies that comparing those who were offered the option to participate in a training program to those who were not correctly identifies the effect of offering people such an option. Thus, the IV estimate using the intention to treat as an instrument correctly estimates the average of the impact of this program on the people who chose to participate. However, this fact does not provide information on the average impact of a training program that was made compulsory for all welfare recipients. To find this out, one would need to set up an experiment with compulsory participation (an interesting outcome would be whether or not people decide to drop out of welfare).

Similar issues also arise in some of the developing-country experiments. For example, the study by Karlan & Zinman (2007) of the effect of access to consumer credit starts from a population of those whose loan applications were rejected by the bank. Then they asked the loan officers to identify a class of marginal rejects from this population and randomly “unrejected” a group of them. However, the loan officers still had discretion and used it to reject approximately half of those who were unrejected. The experiment identifies the effect of this extra credit on the population of those who remained unrejected: It appears to have raised the likelihood that the person remains employed as well as their incomes. However, while this experiment provides (very valuable) evidence that consumer credit may be good for some people, given the unusual nature of the treated population (the twice unrejected), some concern remains that the estimate of the effect of getting a loan in this sample is not representative of this experience for those who are accepted for the loan, or for those who are definitely rejected.

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. However, such a concern is not always valid, as randomization can be used precisely to learn about selection issues. As we discuss above, several studies have been conducted in which the randomization is specifically designed to measure the selection effect, which would be difficult to do in any other way (Karlan & Zinman 2005, Ashraf et al. 2007, Cohen & Dupas 2007). To learn more about selection, Cohen & Dupas (2007) collected hemoglobin levels of women who purchased bed nets at different prices. They were interested in whether women who obtain nets only when they are offered for 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 are specifically collected to study this effect. For example, one important outcome of interest in Ashraf et al. (2006) regards who takes up a self-control device that helps people save.

In other cases, take-up is not an issue because the treatment is in the nature of a pure gift, unlike the offer of training, for example, which is worthless unless someone is also prepared to put in the time. For example, de Mel et al. (2008) studied the effect of offering

each firm in their Sri Lankan sample approximately \$200 in the form of additional capital. They found a large impact on the revenue of the firm, equivalent to a 5–7% return on capital. McKenzie & Woodruff (2008) repeated the same experiment in Mexico and found larger returns (20–35%). In both these cases, the fact that the target firms were small was crucial: The size of the grant ensured that almost everyone was interested in participating in the program (even with gifts, there is always some cost of participation) and allowed such a small gift (which is all they could afford) to have a discernable impact.

However, sometimes even a gift may be refused, as we discovered to our surprise while working with the microfinance institution Bandhan to evaluate its programs designed to help the ultra poor (one of the several evaluations of this program mentioned above) (A. Banerjee, R. Chattopadhyay, E. Duflo & J.M. Shapiro, unpublished results). Under the Bandhan program, villagers who are too poor to be brought into the microfinance net are identified through participatory resource assessments 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 (but with the expectation that they will take care of the asset and there will be 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 whether they simply sell the asset and exhaust 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 back—others 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.

### 3.3. Randomization Issues

The Bandhan study offers an example of randomization bias, i.e., Hawthorne effect: Being part of an experiment (and being monitored) influences behavior. The fact that these villagers were not accustomed to having a private organization go around and give away assets certainly contributed to the problem. However, Bandhan may not have put in the kind of public relations effort to inform the villagers about why the program was being conducted, 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 the potential of creating bad feeling due to the randomization. Location-level randomization is justified by budget and administrative capacity, which is precisely why the organizations often agree to randomize at that level. 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 receive certain programs whereas others do not, and when an NGO serves only some villages, they see it as a part of the organization's overall strategy. When the control areas are given the explanation that the program has enough budget for a certain number of schools only, they typically agree that a lottery is a fair way to allocate those limited resources. They are often used to such arbitrariness and so randomization appears both transparent and legitimate.

One issue with the explicit acknowledgment of randomization as a fair way to allocate the program is that implementers may 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 such 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 program, where control villages knew that they would 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, 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 randomization was involved. This happens more often when randomization takes place at the individual level (though some individual-level randomizations are carried out by public lottery). In such cases, the selected beneficiaries are informed only that, for example, they received a loan for which they had applied (Karlan & Zinman 2007) or that the bank had decided to lower the interest rate (Karlan & Zinman 2005).

### 3.4. Equilibrium Effects

A related issue is what is usually (slightly confusingly) called general equilibrium effects (we prefer the term equilibrium effects because general equilibrium is essentially a multi-market concept). Program effects found in a small study may not generalize what will happen when the program is scaled up nationwide (Heckman et al. 1999, Abbring & Heckman 2008). Consider, for example, what would happen if we tried 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: crowding in the private schools (and potentially a collapse of public schools) and a decline in the returns to education because of increased supply. Both challenges could prompt the experimental evidence to overstate the returns to a nationwide vouchers program.

This phenomenon of equilibrium effects poses a problem that has no perfect solution. Fortunately, in many instances, this phenomenon does not present itself. For example, if we want to determine 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 et al. 2008), the experimental method poses no problem. The fact that immunizing the entire district would not require that many extra nurses helps us here because we can assume that the price of nurses would not increase much, if at all. In another example, although learning that those who received vouchers in Colombia do better in terms of both educational and life outcomes is useful (see Angrist et al. 2002, 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, this concern is precisely one of the reasons why the government may want to carry out such a program. A similar issue arises in the evaluation of training programs. For example, Attanasio et al. (2008b) used randomized allocation of applicants to the job

training program *Jovenes in Action* in Colombia to evaluate its impact. They found that the program had a large effect on the employment rate after graduation. However, because the training program also offered placement aid, it may have also helped the trainees “jump the queue” to get a job. Although part of the impact of a relatively small program, this effect could be muted or entirely disappear once all the youth in a city benefit from the 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—but rather, micro studies using large-scale policy shifts. Even though they are typically not randomized, micro studies still offer the opportunity to be careful about causality issues as well as equilibrium effects, many of which become internalized. A good example of this kind of research is the work of Hsieh & 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 & Angrist (2001) and Duflo (2004a).

It is possible to check if the results from quasi-experimental area-level study 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 those found in experiments. If we find the opposite, we may start worrying about whether the larger study is reliable or representative. In this sense, experiments and nonexperimental studies may be complements rather than substitutes.

Another approach is to try to estimate directly the size of the equilibrium effect using the experimental method. In ongoing research, M. Kremer & K. Muralidharan (unpublished results) study the effect of a vouchers program using a double randomization: They randomize villages where the vouchers are distributed as well as the individuals who receive 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. This approach deals with only one level of equilibration—people can move to the village from outside and leave the village to find work, in which case estimating what is happening to the supply of education rather than to the price of skills may be a better approach—but it is clearly an important start.

An alternative approach is to combine the results from different experiments by using one experiment (or more plausibly, quasi-experiment) to estimate the elasticity of demand for skills, another to estimate the supply of quality teaching, and a third to estimate how much vouchers contribute to skill building. This style of work requires taking a more structural approach because we need to identify the relevant parameters. It, however, has the potential to bridge the gap between the micro and macro worlds and addresses the criticism that experiments may get the right answer to minor questions but fail to address the “big” questions of interest [as seen in some of the comments to Banerjee’s piece in the Boston review, for example, published in Banerjee (2007)]. As we discuss above, experiments can help us estimate economic parameters (such as the returns to capital for small firms, the labor supply elasticity, individual returns to education, etc.), which can then be used in combination with microfounded equilibrium models [Heckman et al. (1999) developed and exposed this method for tuition policy]. There is a small but growing

literature in development economics, associated in particular with Robert Townsend and his collaborators, that attempts to integrate microestimates into calibration of growth models with credit constraints.<sup>6</sup> Clearly, much work remains to be done in this area.

### 3.5. Heterogeneity in Treatment Effects

Most evaluations of social programs focus exclusively on the mean impact. In fact, one of the advantages of experimental results is their simplicity: They are easy to interpret because all you need to do is compare means, a fact that may encourage policymakers to take the results more seriously (see, e.g., Duflo 2004b, Duflo & Kremer 2004). However, as Heckman et al. (1997b) point out, the mean treatment effect may not be what the policymaker wants to know: Exclusive focus on the mean is valid only under specific assumptions about the form of the social welfare function. Moreover, from the point of view of the overall intellectual project, restricting the analysis to the naïve comparison of means does not make sense.

Unfortunately, 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 the need for additional assumptions (Heckman 1992). Of course, we can always compare the entire distribution of outcomes in treatment with that in control: Tests have been developed to measure the equality of distributions as well as stochastic dominance (see Abadie 2002). For example, Banerjee et al. (2007) showed that the distribution of test scores among the students who study in schools that received a *balsakhi* (i.e., “child’s friend” or tutor) first-order stochastically dominates that of the treatment group, and most of the gains are seen at the bottom. This finding is important because, in the program classrooms, the children at the bottom were pulled out and given remedial teaching, whereas those at the top remained in the classroom. We would therefore expect different effects on the two groups, and justifying the program would be difficult if it only helps those at the top. Duflo et al. (2007) also looked at how the camera-based teacher incentive program discussed above affects the entire distribution of absence among teachers, and they found first-order stochastic dominance. However, comparing these distributions does not inform us about the distribution of the treatment effect per se (because 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 & 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. The planner may care about the percentage of people who benefit from a treatment, which is not identified by experiments (or any other evaluation method) without further assumption. To see the issue in its starkest form, consider the following example: There is a population of three people, and we know their

---

<sup>6</sup>We discuss this literature in Banerjee & Duflo (2005). See also Banerjee (2008) for a detailed response to the argument that researchers should give up microestimates because the only thing that matters is growth and that the use of aggregate data is the only way to estimate what drives growth.

potential outcomes both with and without treatment. If not treated, Mr. A's potential outcome is 1, Mr. B's is 2, and Mr. C's is 3. If treated, Mr. A's potential outcome is 2, Mr. B's outcome is 3, and Mr. C's outcome is  $-4$ . What should we think of this program? 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 nontreated 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, because the treatment benefits a majority and the policymaker may care about the greatest good for 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 the potential value of identifying the set of people (from an *ex ante* undifferentiated group) who moved up or down owing to the treatment, a problem arises: Extracting this information from the distribution of outcomes in treatment and in control is impossible. The problem here is logical and not a function of the 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 et al. (1997b) show that the introduction of additional behavioral assumptions (in effect, modeling the decision to participate as a function of the potential outcomes under treatment and nontreatment) allows estimation of precise bounds on features of the distribution of the treatment effect. Abbring & Heckman (2008) provide a detailed treatment of methods to estimate the distribution of treatment effects. These techniques also apply in nonexperimental settings, but the authors point out that they may be particularly useful with experimental data both because one “can abstract from selection problems that plague non-experimental data” and because the experimental setting guarantees balance in the support of the observable variables, something on which the techniques rely.

Our view is that experimental research would benefit by engaging more with this body of research. Reporting additional “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 few assumptions, one can know 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 may need to invoke assumptions that make random assignment less important, cannot possibly be counted in favor of methods not based on random assignment.

Moreover, a lot of the heterogeneity that features prominently in people's objective functions (as opposed to heterogeneity that drives economic outcomes) is not about unobserved differences in people's characteristics, but about potentially observable differences. For example, in the *balsakhi* experiment (Banerjee et al. 2007), not only did we observe that the distribution of test scores in treatment first-order stochastically dominated that in control, but 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 mattered, but we could know this only 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. Crump et al. (2009) developed two nonparametric tests of whether heterogeneity is present in treatment effects: one to

determine whether the treatment effect is zero for any subpopulation (defined by covariates) and another for whether the treatment effect is the same for all subpopulations (defined by covariates). Heckman et al. (2006) and Heckman & Vytlacil (2008a,b) discuss the implication of treatment heterogeneity in terms of both observable and nonobservables.

In addition, treatment effects can be estimated for different subgroups. One difficulty here is that if the subgroups are determined *ex post*, there is a danger of “specification searching,” where researchers and policymakers choose *ex post* to emphasize the program’s impact on one particular subgroup. As in the application by Heckman et al. (1997b), theory can help by telling us what to expect. Specifying *ex ante* the outcomes to be observed and what we expect from them (as is encouraged in the medical literature) is another possibility. Should we still want to try to learn from (possibly interesting, but *ex ante* unexpected) differences in the treatment effect, replication can help: When a second experiment is run, it can be explicitly set up to test this newly generated hypothesis. For example, both Karlan & Zinman (2007) and de Mel et al. (2009) found different results for men and women. These differences were not expected, and they are difficult to reconcile. But once the study is replicated elsewhere, these differences can form the basis of a new set of hypotheses to be tested [see Duflo (2007) for a more detailed discussion of these and other design issues].

Finally, some recent literature (Manski 2000, 2002, 2004; Dehejia 2005; Hirano & Porter 2005) seeks to make all of this less ad hoc. The authors 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 who is trying to decide not only 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 only with expected income gain, but also 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.

### 3.6. Relationship with Structural Estimation

Most of the early experimental literature focused on reduced-form estimates of the program effect. However, there is no reason not to use that data to extract structural parameters wherever possible. Although doing so will require us to make more assumptions, the structural estimates can be used to cross-check the reduced-form results (for example, 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 is obtained from the reduced form.

Early examples of this method in development include Attanasio et al. (2001) and Todd & Wolpin (2006), both of which use PROGRESA data. Attanasio et al. (2001) were

interested in evaluating the program's impact while allowing, for example, for anticipation effects in the control (which cannot be done without making some additional assumptions). They found no evidence of anticipation effects. Todd & Wolpin (2006) wanted to use the experiment as a way to validate the structural model: They estimated a structural model outside the treated sample and checked that the model correctly predicts the impact of the treatment. Another example of the potential of marrying experiments and structural estimation is Duflo et al. (2007). After reporting the reduced-form results, the paper exploits the nonlinear aspect of Seva Mandir's 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 teachers' absences 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 per day by the end of the month, thereby giving up a stochastic outside option of not going to school this day. Yet, this model also gives rise to interesting estimation problems once we try to introduce heterogeneity and serial correlation in the shock received by the teacher on the outside option in a realistic way. As with Todd & Wolpin (2006), this paper then compares the predictions of various models to both the control and a "natural experiment" when Seva Mandir changed their payment rules (after the experiment period was over). This exercise shows that accounting for heterogeneity and serial correlation is important because 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 exploit even further the complementarity between structural estimation and experiments. As mentioned above, one advantage of experiments is their flexibility with respect to data collection and the choice of treatments (within the limits of ethical rule and human subject reviews and what partners are willing to and capable of implementing): It should be possible to design the experiment to facilitate structural estimation by ensuring that the experiment includes sources of variation that would help researchers identify the necessary parameters and collect the appropriate data. Experiments in development economics increasingly involve complex designs and many treatment groups, demonstrating the feasibility of introducing variation that could help identify structural parameters of interest. 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 have yet to see examples of this kind of work: The examples we discuss exploit *ex post* variation in the way the program is implemented, rather than introducing it on purpose.

### 3.7. Relation to Theory

We have argued that experiments can be and have been useful for testing theories [see Banerjee (2005) and Duflo et al. (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. This understanding has prompted us to rethink some basic elements of demand theory.

A number of independent randomized studies on the demand for so-called health protection products consistently found that the price elasticity of demand around zero is

huge. In Kenya, Kremer & Miguel (2007) found that raising the price of deworming drugs from 0 to 30 cents per child reduced the fraction of children taking the drug from 75% to 19%. Also in Kenya, Cohen & Dupas (2007) found that raising the price of insecticide-treated bed nets from 0 to 60 cents reduces the fraction of those who buy the nets by 60%. In Zambia, raising the price of water disinfectant from 9 to 24 cents reduces the fraction of people who take up the offer by 30% (Ashraf et al. 2007). Similar large responses are also found with small subsidies: In India, Banerjee et al. (2008) found that offering mothers one kilogram of dried beans (worth approximately 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%. Most remarkably, a reward of 10 cents got 20% more people in Malawi to pick up the results of their HIV test (Thornton 2007).

Reviewing this evidence (and several papers on education with similar conclusions), Kremer & Holla (2008) 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 either conventionally rational economic agents decide to get an HIV test (knowing their status could prolong their life and that of others) or they decide against getting it (the test may be extremely stressful and shameful). What is more difficult to predict is that so many of them change their minds, for a mere 10 cents, about something that has a good chance of entirely transforming their lives.

Kremer & Holla (2008) suggest that this pattern of demand is more consistent with a model in which people actually want the product but are procrastinating; it is tempting to delay paying the cost given that the benefits are in the future. However, if people really want to buy bed nets or want to 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. In terms of financial products, the (experimental) evidence argues against their being that naïve. Ashraf et al. (2006) found that those who show particularly hyperbolic preferences are also particularly keen to acquire commitment devices to lock in their savings, indicating a high degree of self-awareness. Duflo et al. (2008c,d) found 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 several months later. Moreover, when given *ex ante* (before the harvest), the choice about when vendors should come to sell fertilizer, almost half the farmers request that the vendors come right after harvest, rather than later when the farmers will need fertilizer, because the farmers know that they will have money after the harvest. Their request for fertilizer to be delivered to them right away suggests that the farmers have enough self-control to keep fertilizer at home and not resell it. This finding further suggests that the theory may extend 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 (2007) for a longer discussion]. Bertrand et al. (2009) provide one striking example that fits no existing economic theory: They found 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 literature, field experiments play the role traditionally played by lab experiments, but perhaps with greater credibility. The goal is better theory, but can theory help us design better experiments and interpret experimental results 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 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 come more closely than the rest? Or is every experimental result *sui generis*?

A useful theory for this purpose is unlikely to be particularly sophisticated. Rather, it would provide a convenient way to reduce dimensionality on the basis of a set of reasonable premises. Banerjee et al. (2009) attempted such an approach for the local public action, but their effort is, at best, partially successful. More work along these lines will be vital.

#### 4. CONCLUSION

We fully concur with Heckman's (1992) main point: To be interesting, experiments need to be ambitious and need to be informed by theory. At this convergence point 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). Economists 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. This process of "creative experimentation," where policymakers and researchers work together to think out of the box and learn from successes and failures, is the most valuable contribution of the recent surge in experimental work in economics.

#### DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

#### ACKNOWLEDGMENTS

We thank Guido Imbens for many helpful conversations and James J. Heckman for detailed comments on an earlier draft.

#### LITERATURE CITED

- Abadie A. 2002. Bootstrap tests for distributional treatment effects in instrumental variables models. *J. Am. Stat. Assoc.* 97(457):284–92
- Abbring JH, Heckman JJ. 2008. Econometrics evaluation of social programs part III: distributional treatment effects, dynamic treatment effects, dynamic discrete choice and general equilibrium policy evaluation. See Heckman & Leamers 2008, pp. 5145–5303
- Abdul Latif Jameel Poverty Action Lab (ALJ-PAL). 2005. *Fighting Poverty: What Works?*, Fall, Issue 1. Cambridge, MA: MIT
- Acemoglu D, Angrist J. 2001. How large are human-capital externalities? Evidence from compulsory schooling laws. In *NBER Macroeconomics Annual 2000*, ed. BS Bernanke, K Rogoff, 15:9–74. Cambridge, MA: NBER

- Angrist J, Bettinger E, Bloom E, Kremer M, King E. 2002. Vouchers for private schooling in Colombia: evidence from randomized natural experiments. *Am. Econ. Rev.* 92(5):1535–58
- Angrist J, Bettinger E, Kremer M. 2006. Long-term educational consequences of secondary school vouchers: evidence from administrative records in Colombia. *Am. Econ. Rev.* 96(3):847–62
- Angrist J, Lang D, Oreopoulos P. 2009. Incentives and services for college achievement: evidence from a randomized trial. *Am. Econ. J. Appl. Econ.* 1:136–63
- Angrist J, Lavy V. 2009. The effect of high school matriculation awards: evidence from group-level randomized trials. *Am. Econ. Rev.* In press (see also NBER Work. Pap. 9389).
- Ashraf N, Berry J, Shapiro JM. 2007. *Can higher prices stimulate product use? Evidence from a field experiment in Zambia*. Work. Pap. 13247, NBER
- Ashraf N, Karlan D, Yin W. 2006. Tying Odysseus to the mast: evidence from a commitment savings product in the Philippines. *Q. J. Econ.* 121(2):635–72
- Attanasio O, Barr A, Camillo J, Genicot G, Meghir C. 2008a. *Group formation and risk pooling in a field experiment*. Mimeogr., Georgetown Univ.
- Attanasio O, Kugler A, Meghir C. 2008b. *Training disadvantaged youth in Latin America: evidence from a randomized trial*. Work. Pap., Inst. Fisc. Stud.
- Attanasio O, Meghir C, Santiago A. 2001. *Education choices in Mexico: using a structural model and a randomized experiment to evaluate Progresá*. Mimeogr., Univ. Coll. Lond.
- Banaji M. 2001. Implicit attitudes can be measured. In *The Nature of Remembering: Essays in Honor of Robert G. Crowder*, ed. HL Roediger III, JS Nairne, I Neath, A Surprenant, pp. 117–50. Washington, DC: Am. Psychol. Assoc.
- Banerjee A. 2002. *The uses of economic theory: against a purely positive interpretation of theoretical results*. Work. Pap. 007, Dep. Econ., MIT
- Banerjee A. 2005. New development economics and the challenge to theory. *Econ. Polit. Wkly.* 40(40):4340–44
- Banerjee A. 2007. *Making Aid Work*. Cambridge, MA: MIT Press
- Banerjee A. 2008. *Big answers for big questions: the presumption of growth policy*. Mimeogr., Dep. Econ., MIT
- Banerjee A, Banerji R, Duflo E, Glennerster R, Khemani S. 2009. *Pitfalls of participatory programs: evidence from a randomized evaluation in education in India*. Work. Pap. 14311, NBER; *Am. Econ. J. Econ. Policy*. Forthcoming
- Banerjee A, Cole S, Duflo E, Linden L. 2007. Remedying education: evidence from two randomized experiments in India. *Q. J. Econ.* 122(3):1235–64
- Banerjee A, Duflo E. 2005. Growth theory through the lens of development economics. In *Handbook of Economic Growth*, ed. S Durlauf, P Aghion, 1A:473–552. Amsterdam: Elsevier Sci. Ltd. North Holl.
- Banerjee A, Duflo E, Glennerster R, Kothari D. 2008. *Improving immunization coverage in rural India: a clustered randomized controlled evaluation of immunization campaigns with and without incentives*. Mimeogr., Dep. Econ., MIT
- Banerjee A, Jacob S, Kremer M, Lanjouw J, Lanjouw P. 2005. *Moving to universal education! Costs and trade offs*. Mimeogr., Dep. Econ., MIT
- Beaman L, Chattopadhyay R, Duflo E, Pande R, Topalova P. 2009. *Powerful women: does exposure reduce bias?* BREAD Work. Pap. 181; Work. Pap. 14198, NBER; *Q. J. Econ.* Forthcoming
- Berry J. 2008. *Child control in education decisions: an evaluation of targeted incentives to learn in India*. Mimeogr., Dep. Econ., MIT
- Bertrand M, Chugh D, Mullainathan S. 2005. Implicit discrimination. *Am. Econ. Rev.* 95(2):94–8
- Bertrand M, Djankov S, Hanna R, Mullainathan S. 2007. Corruption in driving licensing process in Delhi. *Q. J. Econ.* 122(4):1639–76
- Bertrand M, Karlan D, Mulainathan S, Zinman J. 2009. What's advertising content worth? Evidence from a consumer credit marketing. *Q. J. Econ.* Forthcoming
- Bjorkman M, Svensson J. 2007. *Power to the people: evidence from a randomized field experiment of a community-based monitoring project in Uganda*. Work. Pap. 6344, CEPR; *Q. J. Econ.* Forthcoming

- Bleakley H. 2007. Disease and development: evidence from hookworm eradication in the American south. *Q. J. Econ.* 122(1):73–117
- Bobonis G, Miguel E, Sharma CP. 2006. Anemia and school participation. *J. Hum. Resour.* 41(4):692–721
- Cohen J, Dupas P. 2007. *Free distribution or cost-sharing? Evidence from a randomized malaria prevention experiment.* Glob. Work. Pap. 14, Brookings Inst.
- Crump R, Hotz J, Imbens G, Mitnik O. 2009. Nonparametric tests for treatment effect heterogeneity. *Rev. Econ. Stat.* In press
- Dehejia R. 2005. Program evaluation as a decision problem. *J. Econom.* 125:141–73
- de Mel S, McKenzie D, Woodruff C. 2008. Returns to capital in microenterprises: evidence from a field experiment. *Q. J. Econ.* 123(4):1329–72
- de Mel S, McKenzie D, Woodruff C. 2009. Are women more credit constrained? Experimental evidence on gender and microenterprise returns. *Am. Econ. J. Appl. Econ.* Forthcoming
- Duflo E. 2004a. The medium run consequences of educational expansion: evidence from a large school construction program in Indonesia. *J. Dev. Econ.* 74(1):163–97
- Duflo E. 2004b. Scaling up and evaluation. In *Accelerating Development*, ed. F Bourguignon, B Pleskovic, pp. 342–67. Washington, DC: World Bank/Oxford Univ. Press
- Duflo E. 2007. Field experiments in development economics. In *Advances in Economic Theory and Econometrics*, ed. R Blundell, W Newey, T Persson; Econ. Soc. Monogr. 42, chpt. 13. Cambridge, UK: Cambridge Univ. Press
- Duflo E, Chattopadhyay R. 2004. Women as policy makers: evidence from a randomized policy experiment in India. *Econometrica* 72(5):1409–43
- Duflo E, Dupas P, Kremer M. 2008a. *Peer effects, pupil teacher ratios, and teacher incentives: evidence from a randomized evaluation in Kenya.* Mimeogr. Dep. Econ., MIT
- Duflo E, Dupas P, Kremer M, Sinei S. 2006. *Education and HIV/AIDS prevention: evidence from a randomized evaluation in western Kenya.* Work. Pap. 402, World Bank Policy Res.
- Duflo E, Hanna R, Ryan S. 2007. *Monitoring works: getting teachers to come to school.* BREAD Work. Pap. 103 (Work. Pap. 11880, NBER)
- Duflo E, Kremer M. 2004. Use of randomization in the evaluation of development effectiveness. In *Evaluating Development Effectiveness*, World Bank Ser. Eval. Dev., Vol. 7, ed. O Feinstein, GK Ingram, GK Pitman, pp. 205–32. New Brunswick, NJ: Transactions
- Duflo E, Kremer M, Glennerster R. 2008b. Using randomization in development economics research: a toolkit. In *Handbook of Development Economics*, Vol. 4, ed. T Schultz, J Strauss, chpt. 15. Amsterdam: Elsevier Sci. Ltd. North Holl.
- Duflo E, Kremer M, Robinson J. 2008c. How high are rates of return to fertilizer? Evidence from field experiments in Kenya. *Am. Econ. Rev. Pap. Proc.* 98(2):482–88
- Duflo E, Kremer M, Robinson J. 2008d. *Why are farmers not using fertilizer? Procrastination and learning in technology adoption.* Mimeogr., Dep. Econ., MIT
- Dupas P. 2007. *Relative risks and the market for sex: teenage pregnancy, HIV, and partner selection in Kenya.* Mimeogr., Univ. Calif. Los Angeles
- Fiszbein A, Schady N, eds. 2009. *Conditional Cash Transfers: Reducing Present and Future Poverty.* Washington, DC: World Bank
- Gine X, Karlan D, Zinman J. 2008. *Put your money where your butt is: a commitment savings account for smoking cessation.* Mimeogr., Dep. Econ., Yale Univ.
- Glewwe P, Ilias N, Kremer M. 2003. *Teacher incentives.* Work. Pap., Dep. Econ., Harvard Univ.
- Glewwe P, Kremer M, Moulin S. 2009. Many children left behind? Textbooks and test scores in Kenya. *Am. Econ. J. Appl. Econ.* 1:112–35
- Glewwe P, Kremer M, Moulin S, Zitzewitz E. 2004. Retrospective vs. prospective analyses of school inputs: the case of flip charts in Kenya. *J. Dev. Econ.* 74(1):251–68
- Heckman JJ. 1992. Randomization and social policy evaluation. In *Evaluating Welfare and Training Programs*, ed. C Manski, I Garfinkel, pp. 201–30. Cambridge, MA: Harvard Univ. Press

- Heckman JJ, Ichimura H, Smith J, Todd P. 1998. Characterizing selection bias using experimental data. *Econometrica* 66:1017–98
- Heckman JJ, Ichimura H, Todd P. 1997a. Matching as an econometric evaluation estimator: evidence from evaluating a job training program. *Rev. Econ. Stud.* 64:605–54
- Heckman J, Leamers E. 2008. *Handbook of Econometrics*, Vol. 6B. Amsterdam: Elsevier Sci. Ltd. North Holl. 1054 pp.
- Heckman JJ, Lochner L, Taber C. 1999. Human capital formation and general equilibrium treatment effects: a study of tax and tuition policy. *Fisc. Stud.* 20(1):25–40
- Heckman JJ, Smith J, Clements N. 1997b. Making the most out of programme evaluations and social experiments: accounting for heterogeneity in programme impacts. *Rev. Econ. Stud.* 64:487–535
- Heckman JJ, Schmierer D, Urzua S. 2010. Testing the correlated random coefficient model. *J. Econom.* Forthcoming
- Heckman JJ, Urzua S, Vytlačil EJ. 2006. Understanding instrumental variables in models with essential heterogeneity. *Rev. Econ. Stat.* 88(3):389–432
- Heckman JJ, Vytlačil EJ. 2008a. Econometrics evaluation of social program part I: using the marginal treatment effect to organize alternative economic estimators to evaluate social programs and to forecast their effect in new environment. See Heckman & Leamers 2008, pp. 4779–874
- Heckman JJ, Vytlačil EJ. 2008b. Econometrics evaluation of social program part II: using the marginal treatment effect to organize alternative economic estimators to evaluate social programs and to forecast their effect in new environment. See Heckman & Leamers 2008, pp. 4875–5144
- Hirano K, Porter J. 2005. Asymptotics for statistical decision rules. *Econometrica* 71(5):1307–38
- Hsieh C-T, Urquiola M. 2006. The effects of generalized school choice on achievement and stratification: evidence from Chile's voucher program. *J. Public Econ.* 90:1477–503
- Imbens G, Angrist J. 1994. Identification and estimation of local average treatment effects. *Econometrica* 61(2):467–76
- Imbens G, Wooldridge JM. 2008. Recent developments in the econometrics of program evaluation. Mimeogr., Dep. Econ., Harvard Univ.; *J. Econ. Lit.* Forthcoming
- Karlan D, Zinman J. 2005. *Observing unobservables: identifying information asymmetries with a consumer credit field experiment*. Work. Pap. 94, Dep. Econ., Yale Univ.
- Karlan D, Zinman J. 2007. *Expanding credit access: using randomized supply decisions to estimate the impacts*. Mimeogr., Dep. Econ., Yale Univ.
- Karlan D, Zinman J. 2008. Credit elasticities in less developed countries: implications for microfinance. *Am. Econ. Rev.* 98(3):1040–68
- Kremer M, Holla A. 2008. *Pricing and access: lessons from randomized evaluation in education and health*. Mimeogr., Dep. Econ., Harvard Univ.
- Kremer M, Miguel E. 2007. The illusion of sustainability. *Q. J. Econ.* 122(3):1007–65
- Kremer M, Miguel E, Thornton R. 2007. *Incentives to learn*. Work. Pap. 10971, NBER; *Rev. Econ. Stat.* Forthcoming
- Manski C. 2000. Identification problems and decisions under ambiguity: empirical analysis of treatment response and normative analysis of treatment choice. *J. Econom.* 95:415–42
- Manski C. 2002. Treatment choice under ambiguity induced by inferential problems. *J. Stat. Plan. Inference.* 105:67–82
- Manski C. 2004. Statistical treatment rules for heterogeneous populations. *Econometrica.* 2(4):1221–46
- McKenzie D, Woodruff C. 2008. Experimental evidence on returns to capital and access to finance in Mexico. *World Bank Econ. Rev.* 22(3):457–82
- Miguel E, Kremer M. 2004. Worms: identifying impacts on education and health in the presence of treatment externalities. *Econometrica.* 72(1):159–217

- Olken B. 2007. Monitoring corruption: evidence from a field experiment in Indonesia. *J. Polit. Econ.* 115(2):200–49
- Rodrik D. 2008. *The new development economics: we shall experiment, but how shall we learn?* Mimeogr., Dep. Econ., Harvard Univ.
- Rubin D. 2006. *Matched Sampling for Causal Effects*. Cambridge, UK: Cambridge Univ. Press
- Topalova P, Duflo E. 2003. *Unappreciated service: performance, perceptions, and women leaders in India*. Mimeogr., MIT
- Thornton R. 2007. The demand for and impact of HIV testing. Evidence from a field experiment. *Am. Econ. Rev.* 98(5):1829–63
- Todd P, Wolpin KI. 2006. Using experimental data to validate a dynamic behavioral model of child schooling: assessing the impact of a school subsidy program in Mexico. *Am. Econ. Rev.* 96(5):1384–417
- World Bank. 2008. *De nouveaux modes de gestion pour accroître les performances de l'enseignement primaire malgache*. Work. Pap., World Bank



# Contents

Some Developments in Economic Theory Since 1940: An Eyewitness Account <i>Kenneth J. Arrow</i> . . . . .	1
School Vouchers and Student Achievement: Recent Evidence and Remaining Questions <i>Cecilia Elena Rouse and Lisa Barrow</i> . . . . .	17
Organizations and Trade <i>Pol Antràs and Esteban Rossi-Hansberg</i> . . . . .	43
The Importance of History for Economic Development <i>Nathan Nunn</i> . . . . .	65
Technological Change and the Wealth of Nations <i>Gino Gancia and Fabrizio Zilibotti</i> . . . . .	93
CEOs <i>Marianne Bertrand</i> . . . . .	121
The Experimental Approach to Development Economics <i>Abhijit V. Banerjee and Esther Duflo</i> . . . . .	151
The Economic Consequences of the International Migration of Labor <i>Gordon H. Hanson</i> . . . . .	179
The State of Macro <i>Olivier Blanchard</i> . . . . .	209
Racial Profiling? Detecting Bias Using Statistical Evidence <i>Nicola Persico</i> . . . . .	229
Power Laws in Economics and Finance <i>Xavier Gabaix</i> . . . . .	255
Housing Supply <i>Joseph Gyourko</i> . . . . .	295

Quantitative Macroeconomics with Heterogeneous Households <i>Jonathan Heathcote, Kjetil Storesletten, and Giovanni L. Violante . . . . .</i>	319
A Behavioral Account of the Labor Market: The Role of Fairness Concerns <i>Ernst Fehr, Lorenz Goette, and Christian Zehnder . . . . .</i>	355
Learning and Equilibrium <i>Drew Fudenberg and David K. Levine . . . . .</i>	385
Learning and Macroeconomics <i>George W. Evans and Seppo Honkapohja . . . . .</i>	421
Sufficient Statistics for Welfare Analysis: A Bridge Between Structural and Reduced-Form Methods <i>Raj Chetty . . . . .</i>	451
Networks and Economic Behavior <i>Matthew O. Jackson . . . . .</i>	489
Improving Education in the Developing World: What Have We Learned from Randomized Evaluations? <i>Michael Kremer and Alaka Holla . . . . .</i>	513
Subjective Probabilities in Household Surveys <i>Michael D. Hurd . . . . .</i>	543
Social Preferences: Some Thoughts from the Field <i>John A. List . . . . .</i>	563

## Errata

An online log of corrections to *Annual Review of Economics* articles may be found at <http://econ.annualreviews.org>