2

The New Development Economics: We Shall Experiment, but How Shall We Learn?

DANI RODRIK

Development economics has long been split between the study of *macro*development (economic growth, international trade, and fiscal/macropolicies) and *micro*development (microfinance, education, health, and other social programs). Even though the central question animating both branches ostensibly is how to achieve sustainable improvements in living standards in poor countries, their concerns and methods have at times diverged so much that they seem at opposite extremes of the economics discipline.

I argue in this chapter that it is now possible to envisage a reunification of the field as these sharp distinctions are eroding in some key respects. Microdevelopment economists have become more interested in policy questions (as opposed to theory and hypothesis testing), and their experimental approach nicely dovetails with some of the current macropolicy trends on the ground. But there is also some basis for pessimism, related to the divergence in empirical methods.

The good news is the substantial convergence in the policy mind-set exhibited by microevaluation enthusiasts, on the one hand, and growth diagnosticians, on the other. The emerging "consensus" revolves not around a specific list of policies, but around how one *does* development policy. In fact, practitioners of this "new" development economics—whether of the "macro" type or "micro" type—tend to

I am grateful to Pranab Bardhan, Tim Besley, Jessica Cohen, Angus Deaton, Pascaline Dupas, Ricardo Hausmann, Asim Khwaja, Sendhil Mullainathan, Mead Over, Lant Pritchett, and Martin Ravallion for their comments on an earlier draft.

be suspicious of claims to ex ante knowledge about what works and what does not work. In their view, the answer lies neither in the Washington Consensus nor in any specific set of initiatives in health or education. Instead it should emerge from a recognition of the *contextual* nature of policy solutions. Relative ignorance calls for an approach that is explicitly *experimental*, and that is carried out using the tools of *diagnostics* and *evaluation*. Old dichotomies between states and markets play little role in this worldview, and *pragmatism* reigns. The proof of the pudding is in the eating: if something works, it is worth doing.

This convergence has remained largely hidden from view because the analytical and empirical tools used by economists at the macro and micro end of things—growth economists versus social policy economists—tend to be quite different. Nevertheless, such a convergence is clearly under way, it is a significant departure from the approaches that dominated thinking about development policy until a decade or so ago, and it represents a significant advance over the previous generation of research.

The bad news is the accentuation of the methodological divergence, which threatens to overshadow the convergence on policy. In particular, the randomized field trials revolution led by researchers in and around the MIT Poverty Action Lab, though greatly enriching the micro end of the field, has created bigger barriers between the two camps.[1] This is not just because randomization is rarely possible with the policies—such as trade, monetary, and fiscal—that macrodevelopment economists study. More important, it is because the stakes are higher with regard to what counts as "admissible" evidence in development. The "randomistas" (as Angus Deaton has called them) tend to think that credible evidence can be generated only with randomized field trials (or when nature cooperates by providing the opportunity of a "natural" experiment).[2] As Abhijit Banerjee puts it: "When we talk of hard evidence, we will therefore have in mind evidence from a randomized experiment, or, failing that, evidence from a true *natural experiment*, in which an accident of history creates a setting that mimics a randomized trial."[3] Randomized field experiments provide "hard" evidence, and by and large only such experiments do. Deprived of randomized (or natural) experiments, macrodevelopment economists would appear to be condemned to second-tier status as peddlers of soft evidence.

So randomizers tend to think real progress is possible only with their kind of evidence. For example, Esther Duflo attributes the periodic shifts in development

1. Banerjee (2007); Duflo (2006); Duflo, Glennerster, and Kremer (2006).

2. Deaton (2007).

3. Banerjee (2007, p. 12). This sentence is preceded by a paragraph that recognizes the weaknesses of inferences from such "hard evidence" (especially with respect to external validity and the feasibility of randomization), and that ends with a much more limited goal: "One would not want to spend a lot of money on an intervention without doing at least one successful randomized trial *if one is possible.*"

policy paradigms and the fact that policy debates never seem to be resolved to the weakness of the evidentiary base to date. Randomization, she argues, provides the way out: "All too often development policy is based on fads, and randomized evaluations could allow it to be based on evidence."[4] Similarly, Banerjee argues that aid should be based on the hard evidence that randomized experiments provide, instead of the wishy-washy evidence from cross-country regressions or case studies.[5] When challenged that substantial progress in economic development has been typically due to economy-wide policy reforms (as in China or India recently), rather than the small-scale interventions in health or education that their experiments focus on, the randomizers respond: "That may well be true, but we have no credible evidence on which of these economy-wide policies work or how countries like China have in fact done it; so we might as well do something in areas we can learn something about."

Actually, it is misleading to think of evidence from randomized evaluations as distinctly "hard" in comparison with other kinds of evidence that development economists generate and rely on. This may seem an odd claim in light of the apparent superiority of evidence from randomized trials. As Banerjee puts it: "The beauty of randomized evaluations is that the results are what they are: we compare the outcome in the treatment with the outcome in the control group, see whether they are different, and if so by how much."[6] Case closed? Well, it depends on what the evidence is needed for and how it will be used. Economists might be interested in how responsive farmers are to price incentives or whether poor educational outcomes are driven in part by a lack of information about relative school performance. Policymakers may want to know what the impacts of a fertilizer subsidy or an informational campaign about school performance are likely to be. In each of these instances, a randomized evaluation can provide some guidance, but it will rarely be decisive. The typical evaluation will have been carried out in a specific locale on a specific group and under specific experimental conditions. Its generalizability to other settings—its "external validity"—is never assured—and it is certainly not established by the evaluation itself.[7]

Generalizability issues can be illustrated by a recent evaluation of an experiment in Western Kenya concerning the distribution of insecticide-treated bed nets to pregnant women.[8] Free distribution, Jessica Cohen and Pascaline Dupas find was vastly more effective than charging a small fee for the bed nets. This would appear to debunk the commonly held view that the valuation and usage of bed net must increase with price—at least in the specific setting in which the experimen was carried out. But do the results extend to other settings in Africa as well? On

4. Duflo (n.d., p. 2).
5. Banerjee (2007).
6. Banerjee (2007, pp. 115–16).
7. See also Basu (2005) for a very useful discussion of the limitations of randomized evaluations.
8. Cohen and Dupas (2007).

can certainly make the case that they do, but the arguments would perforce be informal ones and convincing to varying degrees. In fact, such arguments are not too different from those in defense of a set of instrumental variables (IVs) employed in a conventional econometrics study with weaker internal validity. Moreover, in the public discussion of the Cohen-Dupas evaluation, opponents of free distribution were able to offer a wealth of reasons as to why these results could not be generalized. The debate on free distribution versus cost sharing was hardly settled. Randomized evaluation did *not* yield hard evidence when it came to the actual policy questions of interest. This should not have been a surprise: the only truly hard evidence that randomized evaluations typically generate relates to questions that are so narrowly limited in scope and application that they are in themselves uninteresting. The "hard evidence" from the randomized evaluation has to be supplemented with lots of soft evidence before it becomes usable.

The real test of any piece of research is the Bayesian one: does the finding change our priors on the issue of interest? Randomized evaluations do pretty well when they are targeted closely at the policy change under consideration, but less so when they require considerable extrapolation.[9] In the latter case, evidence from randomized field experiments need not be more informative than other types of evidence that may have less airtight causal identification but are stronger on external validity (because of broader geographical or temporal coverage). In practice, internal validity—just like external validity—is not an either-or matter; some studies do better on this score than others and thus deserve more attention. But this preference has to be tempered with a consideration also of external validity. The bottom line is that randomized evaluations do not deserve monopoly rights—or even necessarily pride of place—in moving one's priors on most of the important questions in development economics.

But this discussion is not meant to be a critique of randomized evaluations, which have indeed greatly enriched the economist's empirical toolkit.[10] It is instead a plea for not letting prevailing methodological differences overshadow the larger convergence. My purpose is to get macrodevelopment economists and microdevelopment economists to see that they have much more in common than they realize. The former are increasingly adopting the policy mind-set of the latter, while the latter skate on thinner ice with their empirical work than is often

9. For example, the study by Bertrand and others (2007) of corruption in the driver's license system in Delhi, India, is of tremendous value to anyone who wants to understand and improve the regime of driver's license allocations in India. However, extrapolating from it to corruption in other types of service delivery or in other countries is extremely difficult and would require considerable care.

10. I do not deal here with the criticism that randomized evaluations typically entail very little theorizing, except insofar as this renders extrapolation to other settings more problematic. Even though this may be a legitimate complaint in practice, I do not think it is a fundamental issue. There is nothing in the nature of randomized trials that precludes either theory testing or more explicit use of theory. As I explain later, good use of experimentation in fact relies on explicit theoretical framing. For a broad discussion of the role of theory versus experimentation, see Kanbur (2005).

thought. I use a specific policy problem—whether insecticide-treated bed nets should be distributed for free or at some nominal fee—as a springboard for examining the different types of evidence, including randomized evaluations (in the vein of Cohen and Dupas), for what I believe is a new paradigm in the making.[11]

A Policy Problem: Should Bed Nets Be Given Out for Free?

It is well known that insecticide-treated bed nets (ITNs) are extremely effective in preventing exposure to malaria. It is also well recognized that ITNs should be subsidized rather than sold at cost: ITNs reduce the number of mosquitoes and the malaria parasites that can be passed on to others, so there are externalities involved on top of the direct income and health poverty impacts. The debate revolves around whether ITNs should be handed out for free or at a positive, if still below-cost, price.

One view, articulated forcefully by Jeffrey Sachs, is that ITNs should be free so as to achieve universal access and have the greatest possible impact on the disease. In this view, it is important to ensure ITNs are used by the community at large, rather than solely by those groups that are typically identified as being at greatest risk (mainly pregnant women and young children) and that are targeted by conventional public health campaigns.[12]

The other view is that free distribution is not cost-effective and sustainable, and that ITNs should be made available at a positive, if still nominal, price. There are several arguments in favor of this "cost sharing."[13] First, it may ensure better targeting insofar as only those who are likely to use the bed nets or those at greater risk will want to pay for them. Second, it may increase usage insofar as people are more likely to value something they have paid for (this is the so-called sunk-cost fallacy). Third, having to pay for a good or service is more likely to make users demand accountability on the part of health care providers. Fourth, cost sharing is more likely to sustain a private delivery mechanism over time (unlike free distribution, which relies on periodic public health campaigns). These are the arguments typically used by social marketing groups, which are particularly active in this area.

Obviously, one cannot choose between these two sets of views on the basis of theory or a priori reasoning. Both are plausible and are likely to be correct for a particular distribution of the underlying structural parameters that determine behavior. How does one gather evidence about the empirical validity of these contrasting viewpoints, which ultimately relates to effectiveness in eradicating malaria? Consider three strategies.

11. In case there is any doubt, I should clarify that I use the Cohen-Dupas study not because of any weaknesses in it, but, quite to the contrary, because it is a particularly well-done evaluation on a question of tremendous interest.

12. Sachs and others (2007).

13. See Over (2008).

Reduced-Form Econometrics

One research approach would be to look at the pattern of correlations across regions and over time between the type of policy or program employed and the malaria outcomes on the ground. So imagine the following regression for sub-Saharan Africa:

$$Y_{it} = \alpha P_{it} + \Sigma_j \beta_j P_{it} X_{it}^j + \Sigma_j \gamma_j X_{it}^j + D_i + D_t + \epsilon_{it}, \quad (2\text{-}1)$$

where Y_{it} stands for the malaria outcome of interest (rates of infection or incidence), P_{it} is an index that captures the nature of policy in place (in particular the extent to which the program relies on free distribution versus cost sharing), X_{it}^j is a set of conditioning variables (income level, population density, demography, other health indicators, and so forth), and D_i and D_t are region- and time-fixed effects. This specification allows policy to interact with background conditions, and it also controls for time trends and time-invariant regional differences.

Subject to the caveats mentioned in the next paragraph, this regression can indicate how effective different program types are, and also how effectiveness varies with the conditioning factors. So the expected impact of changing policy from P to P' in a country where the background conditions are given by X^j is simply $\hat{\alpha}(P' - P) + \Sigma_j \hat{\beta}_j X_{it}^j (P' - P)$, where $\hat{\alpha}$ and $\hat{\beta}_j$ are the estimated parameters from the foregoing regression.

The problems in this research design are many. First of all, it is difficult to specify and include all the background conditions that influence or may be correlated with policy effectiveness. That implies that the researcher may have to contend with various sources of omitted-variable bias. In addition, there may not be enough variation over time, so that equation 2-1 may need to be estimated as a pure cross section:

$$Y_i = \alpha P_i + \Sigma_j \beta_j P_i X_i^j + \Sigma_j \gamma_j X_i^j + \epsilon_i. \quad (2\text{-}2)$$

Since one cannot control for time-invariant regional unobservables in this specification, any potential problem of omitted-variables bias becomes that much more severe.

A second problem is how to code and create a quantitative index for the type of policy in place in different regions or countries. Cost-sharing strategies come in many guises, and in any case, few programs will be of the pure free-distribution or cost-sharing types. In addition, one must take into account other aspects of the program as well: how extensive, how well administered, and how well funded it is, and so on.

Most important, any regression of this type will be open to the criticism that the right-hand-side variables, P in particular, are not exogenous, rendering identification of a truly causal effect problematic. Identification requires that P and the error term ϵ be uncorrelated, which is a demanding test. The most obvious source of

bias in this connection is that the programs may have been selected *in response* to the type of malaria challenge being faced in each region. If a government knows or anticipates that free distribution will be more effective, it will use that type of program instead of the other. This is called the program placement effect in the microeconometric literature and wreaks havoc with all cross-sectional econometric work. More generally, interpretation of the coefficients α and β_j is always problematic in view of the fact that programs are not randomly assigned: they are selected for some reason. Any pattern of correlation desired can be generated by specifying those reasons and their cross-sectional variation appropriately.[14]

In practice, an empirical exercise of this sort is likely to generate a conversation and debate between those who find the results credible (the authors and their supporters) and those who have doubts. "You have measured policies very badly," the critics will say. "But here is an alternative measure with greater detail, and it makes little difference to the results," the authors will respond. "Policies are endogenous and respond to malaria outcomes," the critics will object. "But look, all these countries selected their programs for reasons that had little to do with what was going on the ground, and if you do not believe that, here is an instrumentation strategy that uses the identity of the main external donor as an instrument for the type of program," the authors will perhaps respond. The debate will go on and on, and some people will come to think that the results have some credibility, while others will remain unconvinced.

In theory, identification is an either/or thing. Either the causal effect is identified, or it is not. But in practice, identification can be more or less credible. If the study is done reasonably well and the authors have convincing answers to the criticisms leveled against it, one can (or should) imagine that one's priors on the policy question at hand would be moved by the results of the exercise. One would have to be a purist of the extreme kind to imagine that *nothing* could be learned from a regression of this type, regardless of the quality of the supporting argumentation.

Qualitative Evidence: Surveys

One drawback of the econometric strategy is that not many countries may yet have experimented with either cost-sharing or free distribution programs. So there may not be much variation in P of the type needed to identify the effects of concern.

A qualitative research strategy, based primarily on interviews, may be a substitute. Suppose a team of researchers travels around Africa to undertake in-depth interviews with health professionals and service providers. It would pose the following type of questions:

14. Rodrik (2005).

1. How important do you think cost is as an impediment to the use of bed nets in your region, compared with other obstacles (such as availability and knowledge about benefits)?

2. How likely do you think it is that people will value and use ITNs more if they actually pay for them?

3. Do you think private channels of supply are more likely to exist if ITNs are sold at a price?

4. What is the best way to get people who are less vulnerable (that is, adult males) to use ITNs?

One can imagine the response to these questions being coded for use in quantitative analysis. But the main purpose of the interviews would not be statistical analysis but taking stock of the state of "local knowledge"—what people closest to the problem think—about the key questions that determine the relative effectiveness of free distribution versus cost sharing. And open-ended questions such as the fourth one can help reveal new solutions that the outsider may not have thought about before.

Economists tend to be wary of qualitative research and of evidence that is based on interviews. But as Gary King, Robert Keohane, and Sidney Verba have argued, good qualitative studies use the same logic of inference as quantitative ones.[15] Needless to say, in this particular instance interviewees have limited knowledge, have their own preconceptions (which may or may not be idiosyncratic), have a stake in the outcome (which may affect the nature of their responses), and will be influenced by the environment in which they operate. But even with these limitations, their responses should shed some light on the effectiveness question. Indeed, it would be surprising if eliciting local information systematically in this manner did not serve to narrow the range of plausible outcomes. Experiential knowledge cannot be dismissed altogether.

Of course, conclusions from such research would naturally be contested. How representative were the interviewees, and can they really be expected to predict accurately the consequences of this or that program? But the relevant question here is not whether the interviews can provide a definitive answer; it is whether they can move the profession's priors. If the authors of the study have thought their methodology through, they will have answers for their critics that at least some will find convincing. Once again, only an extreme purist would deny that there is potential for learning from this kind of effort. The scientific method can be applied to qualitative as well as quantitative evidence.

Randomized Field Evaluation

Finally, consider a field experiment that randomizes across recipients as to whether they get ITNs for free or at a (subsidized) price. This provides a way to

15. King, Keohane, and Verba (1994).

look directly for any differential effects in uptake and usage and is exactly the method employed by Cohen and Dupas in Western Kenya.[16] Working with twenty prenatal clinics to offer ITNs at varying prices, they divided the clinics randomly into five groups of four, with four of the groups offering the ITNs at a (single) price ranging from $0 to $0.60 per ITN and the fifth serving as the control. They then measured the uptake of ITNs from the clinics and also spot-checked for usage (whether the nets were hanging on beds or not). In addition, they checked the hemoglobin levels (anemia rates) of women getting ITNs to see if cost sharing does a better job of selecting women at greater risk for malaria.

The results were for the most part unambiguous and quite striking. Cost sharing significantly reduced the number of ITNs that ended up in the hands of recipients without increasing actual usage among those who did receive the bed nets. Furthermore, there was no evidence of selection benefits from cost sharing: women who paid a positive price were no sicker than women in the control group. Under reasonable assumptions on private and social benefits, Cohen and Dupas show that free distribution is more cost-effective than cost sharing: the benefits of greater use more than offset additional budgetary costs.

My initial reaction to this study was that it settled the question once and for all. Free distribution is the way to go.[17] However, further reflection and reading on the topic made clear that I had overreached.[18] One can have genuine doubts as to the extent to which the Cohen-Dupas results can be generalized. As the advocates of cost sharing were quick to point out, the setting for this study was special in a number of respects.[19]

First, the area in Western Kenya where the experiment was carried out had been blanketed by social marketers for a number of years, with as many as half a million bed nets already distributed. There is reason to believe that the value of bed net use was already well understood. In other words, the experiment may have benefited from the earlier demand promotion activities of the social marketers.

Second, the experiment was narrowly targeted at pregnant women making visits to prenatal clinics. In other words, the recipients were a subgroup at high risk for malaria and had revealed themselves to be willing to engage with public health services. Moreover, these women were provided with information about malaria risks. The mass-distribution argument of Sachs, by contrast, is based on free distribution to the population at large.

Third, the experiment took care of supplying ITNs to the clinics, thus isolating the supply side from the demand side of the problem. Therefore the experi-

16. Cohen and Dupas (2007).

17. Hence the title of my blog entry summarizing the paper: "Jeff Sachs Vindicated." See http://rodrik.typepad.com/dani_rodriks_weblog/2008/01/jeff-sachs-vind.html.

18. Stimulated in part by comments on my blog post, mentioned in note 17.

19. See Mead Over, "Sachs Not Vindicated" (http://blogs.cgdev.org/globalhealth/2008/01/sachs_not_vindicated.php).

ment did not test the social marketers' claim that some degree of cost sharing is important to establish sustainable supply channels at the retail level.

Fourth, the difference between the subsidized price and zero was perhaps too small to trigger the "sunk-cost fallacy." Therefore, one should not necessarily rule it out in other settings.

The conclusion that cost-sharing advocates would like readers to draw is this: believe the results for Western Kenya at this particular juncture, but do not expect them to hold in other settings with other background conditions.

In terms of the regression framework discussed previously, what the randomized field experiment estimates is not the α and β_j separately, but the composite term $\alpha + \beta_j \Sigma_j X_i^j$, which also depends on the background conditions X_i^j. It identifies, quite accurately, the effect of policy P under one realization of X_i^j—but gives no way of parsing the manner in which those background conditions have affected the outcome and therefore does not allow one to extrapolate to other settings. That is why it is fair game to question the generalizability of the results.[20]

Now, I suspect that Cohen and Dupas (and Sachs) would have some good arguments as to why these objections to the generalizability of the field experiment results are overdrawn and why the results are likely to hold up in other settings as well.[21] And I suspect that the critics would stand their ground in turn. The key point, however, is that the randomized field evaluation cannot settle the larger policy question that motivated it. It is no different in that respect than the other two research strategies discussed earlier in the chapter. Despite the clean identification provided by the randomized field experiment, those who believe they have learned something general about free distribution have to resort to credibility-enhancing arguments that feel rather similar to those that practitioners of cross-section econometrics and qualitative studies have to resort to—although the effort will now be

20. Deaton (2007, pp. 60–61) puts it thus in his comments on Banerjee (2007): "Take Banerjee's example of flip charts. The effectiveness of flip charts clearly depends on many things, of which the skill of the teacher and the age, background, and previous training of the children are only the most obvious. So a trial from a group of Kenyan schools gives us the average effectiveness of flip charts in the experimental schools relative to the control schools for an area in western Kenya, at a specific time, for specific teachers, and for specific pupils. It is far from clear that this evidence is useful outside of that situation. This qualification also holds for the much more serious case of worms, where the rate of reinfection depends on whether children wear shoes and whether they have access to toilets. The results of one experiment in Kenya (in which there was in fact no randomization, only selection based on alphabetical order) hardly prove that deworming is always the cheapest way to get kids into school, as Banerjee suggests." Or as Mookherjee (2005) complains more generally about development microeconometrics: "A well-executed paper goes into a particular phenomenon in a particular location in considerable depth, data permitting. The research is consequently increasingly microscopic in character. We have very little sense of the value of what we have learned for any specific location to other locations." See also Ravallion (2008a) for a critique of randomized evaluations. Deaton's (2009) critique in his Keynes lectures appeared too late to be reflected in the present chapter.

21. For example, the argument that the results may have been contaminated by the prior presence of social marketing is irrelevant if one wants to extend free distribution to other areas of Kenya or Africa where social marketers have also been active.

directed at convincing critics about the generalizability of their results and not about identification or relevance. No, Western Kenya is not really that different from other settings. No, there was ample opportunity in the research design for sunk-cost effects to operate. No, prior exposure to social marketing could not have made a big difference. And so on. If these arguments are perceived as credible to outsiders like me with little stake in the outcome, it will (and should) move one's priors. But it will do no more than that.

Discussion

I have hardly scratched the surface of possible research strategies. One can add various other regression-based approaches such as structural econometrics or regression discontinuity. One can also think of additional qualitative strategies, such as the structured case-study approach. The point is not to be exhaustive but to illustrate that different styles have different strengths and weaknesses. Cross-sectional and panel regressions can have broad coverage and can control for at least some of the background conditions explicitly. Interviews and other qualitative approaches can be carried out in a more open-ended manner, allowing unanticipated new information to play a role. Randomized evaluations can nail down identification within the confines of the experiment.

In the technical jargon, the research strategies I have described have different degrees of internal and external validity. Internal validity relates to the quality of causal identification: Has the study credibly demonstrated a causal link between the policy or treatment in question and the outcome of interest? External validity has to do with generalizability: Are these results valid also for the broader population for which the policy or treatment is being considered? Sound inference requires *both*.

Randomized evaluations are strong on internal validity but produce results that can be contested on external validity grounds—as I illustrated with the malaria experiment. By contrast, the standard econometric and qualitative approaches are weaker on internal validity—but conditional on credible identification, they have fewer problems of external validity. (In the malaria illustration, they cover all or most of Africa as a whole, and they may also have a tempora dimension.)

Some advocates of randomized evaluations would argue that internal validity trumps all else, that there is no point in worrying about generalizability until a causal relationship is demonstrated clearly at least once.[22] Identification is an either/or matter: an effect is either clearly demonstrated or it is not. So nothing other than randomized trials (or perhaps some natural experiments) can possibly help reveal a truly causal effect. As for external validity, it can best be established through repeated replication of field experiments in different settings. In any case

22. For the canonical statement of this position in social psychology, see Campbell and Stanley (1963

one should proceed lexicographically: conduct randomized field experiments and fret about external validity later.

But does this make sense from a decision-theoretic standpoint? Suppose a policymaker needs to figure out which strategy to adopt—*now*. Or a journal editor has to decide whether a piece of research is sufficiently well done and interesting enough to merit publication. In both cases, the relevant point is whether the research *changes the priors on the question of interest.* This means the internal and external validity tests must be applied simultaneously. Identification alone is inadequate, unless there is strong enough reason to believe that the causal effects can be generalized to the broader population of interest. A study lacking internal validity is surely worthless; but a study lacking external validity is almost worthless too. After all, one is not interested in a result that solely applies to pregnant women visiting prenatal clinics in Western Kenya during a period of several months in 2007 and facing a particular schedule of fees. One is interested in whether the results say anything about the respective advantages of free distribution and cost sharing *in general, or in a specific setting that differs from that of the evaluation.*[23]

This is also in line with the revealed preference of the economics profession, which is to think of identification as gradations rather than as binary. Some identification strategies are viewed as more credible than others, and standards regarding what is credible change over time. In practice, internal validity is a matter of degree, just as external validity is. The implication is that the information content of these different kinds of studies cannot be rank-ordered on an a priori basis. The weights that should be put in the Bayesian updating process on (a) randomized evaluations and (b) other types of evidence must both lie strictly between 0 and 1, unless the nonrandomized evidence has no claim to internal validity at all. Moreover, the respective magnitude of these weights cannot be determined on the basis of a priori reasoning (except again in limiting cases). One may well be swayed more by a study that is less than airtight on internal validity but strong on external validity than by a study with strong internal validity but unclear external validity.

23. This is how Banerjee (2005) discusses a similar problem: "If our only really reliable evidence was from India but we were interested in what might happen in Kenya, it probably does make sense to look at the available (low quality) evidence from East Africa. Moreover, if the two types of evidence disagree, we might even decide to put a substantial amount of weight on the less reliable evidence, if it turns out that it fits better with our prior beliefs. Nevertheless, there remains an essential asymmetry between the two: The well-identified regression does give us the 'correct' estimate for at least one population, while the other may not be right for anyone. For this reason, even if we have many low quality regressions that say the same thing, there is no sense in which the high quality evidence becomes irrelevant—after all, the same source of bias could be afflicting all the low quality results. The evidence remains anchored by that one high quality result." I am not sure what the last sentence means, but I agree with the rest, which seems to grant the point that in general both types of evidence should receive positive weight. I am certainly not arguing that "the high-quality evidence" from the randomized

Of course, practitioners of randomized field evaluations do recognize problems of external validity. Duflo and her colleagues in particular provide an excellent and comprehensive discussion of external validity pitfalls in randomized trials.[24] As Duflo puts it: "Even if the choice of the comparison and treatment groups ensures the internal validity of estimates, any method of evaluation is subject to problems with external validity due to the specific circumstances of implementation. That is, the results may not be able to be generalized to other contexts."[25] What is less often recognized is that some methods of evaluation *may* have fewer problems of external validity because they allow greater coverage over time and space of the relevant population. Advocates of randomization easily slip into language that portrays experimental evidence as "hard," overlooking the fact that theirs is as "soft" as other types of evidence when it comes to the real questions at hand.

Consider Banerjee's complaint that the World Bank's sourcebook on empowerment and poverty reduction has only one recommendation based on a randomized trial (school vouchers, subjected to randomized evaluation in Colombia).[26] As for the recommendation on legal reform, he says "the available evidence, which comes from comparing the more law-abiding countries with the rest, is too tangled to warrant such a confident recommendation."[27] He faults the bank both for not showing more enthusiasm for programs like vouchers (for which there is a study with good internal validity) and for endorsing strategies like legal reform (for which there are many studies that do more poorly on internal validity). "What is striking about the list of strategies offered by the World Bank's sourcebook," Banerjee writes, "is the lack of distinction made between strategies based on the *hard* evidence provided by randomized trials or natural experiments and the rest."[28] But of course the experimental evidence from Colombia is equally problematic when it comes to *generalizability* to other countries. How would the results change if, as would be necessary, one altered the target population (children of secondary school age in Colombia's low-income neighborhoods)? Or the environment in which the experiment was conducted (for example, the availability and quality of nearby private educational facilities)? Or some details of the program (for example the share of private school costs covered by the voucher)?[29] No one knows. So it is not at all clear that the priors on the relevant policy question—what strategies are worth pursuing to empower the poor and reduce

24. Duflo and others (2006).
25. Duflo (n.d., p. 27).
26. Banerjee (2007).
27. Banerjee (2007, p. 14).
28. Banerjee (2007, p. 13), emphasis added.
29. The study in question is Angrist and others (2002). The authors conclude, cautiously: "Our findings suggest that demand-side programs like PACES can be a cost-effective way to increase education attainment and academic achievement, at least in countries like Colombia with a weak public school infrastructure and a well-developed private-education sector" (p. 1556).

poverty across the globe—should be moved more by the Colombia study than by the multitude of cross-national studies on legal institutions. The right way to present this would have been to recognize that both types of evidence have strengths and weaknesses when it comes to informing policymakers about the questions they care about.

The need to demonstrate credible identification is well understood in empirical economics today. When I was a young assistant professor, one could still publish econometric results in top journals with nary a word on the endogeneity of regressors. If one went so far as to instrument for patently endogenous variables, it was often enough to state that one was doing IV, with the list of instruments tacked into a footnote at the bottom of a table. No more. A large chunk of the typical empirical—but nonexperimental—article today is devoted to discussing issues having to do with endogeneity, omitted variables, and measurement error. The identification strategy is made explicit and is often at the core of the research. Robustness issues take a whole separate section. Possible objections are anticipated and counterarguments are advanced. In other words, considerable effort is devoted to convincing the reader of the internal validity of the study.

By contrast, the typical study based on a randomized field experiment says very little about external validity. If there are some speculations about the background conditions that may have influenced the outcomes and that do or do not exist elsewhere, they are offered in passing and are not central to the flow of the argument. Most important, the typical field experiment makes no claims about the generalizability of the results—even though without generalizability a field experiment is of little interest, as I have just argued. But little is said to warn the reader against generalizing, either.[30] And since the title, summary, motivation, and conclusions of the study typically revolve around the *general* policy question, careless readers may well walk away from the study thinking that they have learned more about the broader policy question of interest than they actually should have.

Interestingly, in medicine, where clinical trials have a long history, external validity is also a major concern, and it is often neglected. The question there is whether the findings of a randomized controlled trial, carried out on a particular set of patients under a specific set of conditions, can be generalized to the population at large. One recent study complains that published studies do a poor job of reporting on external validity, and that "researchers, funding agencies, ethics committees, the pharmaceutical industry, medical journals, and governmental regulators alike all neglect external validity, leaving clinicians to make judgments."[31] The long list of evidence adduced in support of this argument makes for

30. See the online draft version of Cohen-Dupas (2007), which contains stronger language in its introduction and conclusions warning against extrapolation to other settings (www.brookings.edu/~/media/Files/rc/papers/2007/12_malaria_cohen/12_malaria_cohen.pdf [May 19, 2008]).

31. Rothwell (2005, p. 82).

Box 2-1. *Neglect of Consideration of External Validity of Randomized Controlled Trials (RCTs) in Medicine*

Research into internal validity of RCTs and systematic reviews far outweighs research into how results should best be used in practice.

Rules governing the performance of trials, such as good clinical practice, do not cover issues of external validity.

Drug-licensing bodies, such as the U.S. Food and Drug Administration, do not require evidence that a drug has a clinically useful treatment effect or a trial population that is representative of routine clinical practice.

Guidance on the design and performance of RCTs from funding agencies, such as that from the U.K. Medical Research Council, makes virtually no mention of issues related to external validity.

Guidance from ethics committees, such as that from the U.K. Department of Health, indicates that clinical research should be internally valid and raises some issues that relate to external validity, but makes no explicit recommendations about the need for results to be generalizable.

Guidelines on the reporting of RCTs and systematic reviews focus mainly on internal validity and give very little space to external validity.

None of the many scores for judging the quality of RCTs address external validity adequately.

There are no accepted guidelines on how external validity of RCTs should be assessed.

Source: Reproduced from Rothwell (2005).

interesting reading in light of the parallels with current practice in economics (see box 2-1). Virtually all of these points have their counterpart in current experimental work in development economics.

One response to the external-validity critique is to say that the solution is to repeat the experiment in other settings, and to do it enough times so that the researcher feels confident in drawing general lessons. Repetition would surely help. But is it the magic bullet? Few randomized evaluations—if any—offer a structural model that describes how the proposed policy will work, if it does, and under what circumstances it will not, if it does not. Absent a full theory that is being put to a test, it is somewhat arbitrary to determine under what different conditions the experiment ought to be repeated. If one does not have a theory of which X_i^j's matter, one cannot know how to vary the background conditions. Moreover, everyone is free to come up with an alternative theory that would enlarge the set of conditioning variables. As Ravallion puts it: "The feasibility of doing a sufficient number of trials—sufficient to span the relevant domain of variation found in reality for a given program, as well as across the range of pol-

icy options—is far from clear. The scale of the randomized trials needed to test even one large national program could well be prohibitive."[32]

But the more practical objection to the repetition solution is that there is very little professional incentive to do so. It is hard to imagine that leading journals will be interested in publishing the results of an identical experiment that differs along one or two dimensions: perhaps it is a different locale, or perhaps the policy varies a bit, but in all other ways, the experiment remains the same. The conditions under which the repetition is most useful for purposes of external validity—repetition under virtually identical conditions, save for one or two differences—are precisely the conditions that will make it unappealing for purposes of professional advancement. It is possible that nongovernmental organizations and governments can step in to provide the replication needed. But these actors have their own interests and stakes in the outcome. Their efforts may be as problematic as those from clinical trials undertaken by the pharmaceutical industry.[33]

Perhaps ironically, other types of studies that have weaker internal validity generate much greater incentive for replication. Here the name of the game is improved identification, and there are ample professional benefits for researchers who come up with a new instrumental variable or a novel identification strategy.

Ultimately, the best way to render randomized field trials more useful is to make a careful consideration of external validity part and parcel of the exercise. It should be incumbent on the authors to convince the reader that the results are reasonably general and also to address circumstances under which they may not be. This is as important as justifying causal identification in other types of empirical work. A discussion of external validity will necessarily remain speculative along many dimensions. But that is its virtue: it will bring to the fore what is in many instances a hidden weakness. And the need to justify external validity ex post may also stimulate better experimental design ex ante. For instance, researchers may make a greater effort to target a population that is "representative," be more explicit about the theoretical foundations of the exercise, and incorporate (at least) some variation in the X's.[34]

The Good News: Convergence in Policy Mind-Sets

For Banerjee, "what is probably the best argument for the experimental approach [is that] it spurs innovation by making it easy to see what works."[35] The premise

32. Ravallion (2008a, p. 19).

33. Rothwell (2005).

34. An excellent example of a field experiment that uses theory to guide the exercise and inform issues of external validity is Jensen and Miller (2008). These authors were interested in the existence of a Giffen good, so they carried out the experiment in a setting that theory suggested is most conducive to locating it (very poor Chinese consumers facing variation in the price of their staple foods, rice or noodles). As a by-product, the analysis clarifies the circumstances under which their result would generalize.

35. Banerjee (2007, p. 122).

is that policy innovation is inherently useful—either because problems may need to be solved through unconventional ways or because different contexts require different solutions. This may be an uncontroversial premise in the domain of social policy, but until recently it ran counter to much thinking in the area of growth. Up until a decade or so ago, macrodevelopment economists thought they had a fairly good idea of what it would take to turn economic performance around in the closed, statist economies of Latin America, Africa, the Middle East, and South Asia. These economies needed to remove trade restrictions, free up prices, privatize state enterprises and parastatals, and run tighter fiscal policies. The list was clear-cut and in need of very little innovation or experimentation, save possibly for evading the political minefields associated with these reforms.

While it would be an exaggeration to say that the previous consensus has totally dissipated, today macrodevelopment economists operate in a very different intellectual environment. Gone is the confidence that they have the correct recipe, or that privatization, stabilization, and liberalization can be implemented in similar ways in different parts of the world.[36] Reform discussions focus on the need to get away from "one-size-fits-all" strategies and on context-specific solutions. The emphasis is on the need for humility, for policy diversity, for selective and modest reforms, and for experimentation. Gobind Nankani, the then vice president of the World Bank who oversaw the effort behind the bank's *Economic Growth in the 1990s: Learning from a Decade of Reform,* writes in the preface of the book: "The central message of this volume is that there is no unique universal set of rules. . . . [W]e need to get away from formulae and the search for elusive 'best practices.'"[37] The recent Spence report on growth encapsulates and reflects many of these changed views.[38]

My own work (with colleagues Ricardo Hausmann, Lant Pritchett, Charles Sabel, and Andrés Velasco) has focused on developing methodologies for designing country-specific growth strategies and on innovations in institutional arrangements for industrial policy.[39] We formulate the underlying problem as one of "growth diagnostics": how to discover the binding constraints on economic growth in a specific setting, and then how to come up with policy solutions that are cognizant of local second-best interactions and political constraints. The detective work consists of postulating a series of hypotheses about the nature of the economy and its underlying growth process (or lack thereof) and checking to see whether the evidence is consistent with the signals one would expect to

36. See World Bank (2005); Rodrik (2006).

37. World Bank (2005, p. xiii).

38. Commission on Growth and Development (2008).

39. On country-specific growth strategies, see Rodrik (2007); Hausmann, Pritchett, and Rodrik (2005); Hausmann, Rodrik, and Velasco (2008). On institutional arrangements for industrial policy, see Rodrik (2008); Hausmann, Rodrik, and Sabel (2007).

observe under those hypotheses. In other words, the approach follows the "
tific method" even though the answers it generates necessarily come with
margins of doubt. Policy design in turn relies less on "best practices" and mc
a combination of experimentation and monitoring.

These ideas may have been new in the growth context, but in fact the
parallel to the thinking reflected in the work of microdevelopment econo
focusing on randomized evaluations. For me, the epiphany occurred duri
executive program we were offering at the Harvard Kennedy School, "
Thinking on Economic Growth and Development." I was sitting in on a di
sion that Banerjee was conducting on the health crisis in Rajasthan and po
responses to it (which had been preceded by an excellent video produce
Banerjee and his colleagues). Over the course of the discussion, it became
that the approach Banerjee was taking the class through was virtually identi
the Hausmann-Rodrik-Velasco (HRV) "diagnostic" approach—albeit in a
different setting. No basic presumption is made about having the answer
poor health outcomes are due to inadequate public spending, say, or ignor
about the value of health). So the researcher conducts surveys and interview
collects information.

The next step is to develop stories about what may account for the trou
Are people not receiving good health care because there are no health cl
nearby? Because they do not think clinics are useful? Because there are "cr
doctors who provide apparently substitute services? Or because nurses and
tors are frequently absent? Each of these stories has implications for the patt
apparent in the surveys and the response people give in the interviews (they th
out different "diagnostic signals," in HRV terminology). If poor people spe
considerable share of their budget on health, for example, it is unlikely that
do not value it sufficiently. This kind of analysis helps narrow the list down
smaller list of real problems ("binding constraints"). Then one gets creative
tries to come up with ways—often quite unconventional—in which to overc
these problems (lentils in exchange for inoculation, cameras in the classroom,
so on). Finally, the researcher subjects these ideas to rigorous evaluations thro
randomized experiments and amends them as required.

This thought process captures fairly well the spirit in which growth diagn
tics exercises are supposed to be carried out as well. What my colleagues and I
begun to advocate for macrodevelopment economists was exactly the same k
of open-minded, open-ended, pragmatic, experimental, and contextual appro
If our ideas seemed (at the time, but perhaps no longer) unorthodox, it
largely because there was already a Washington Consensus to contend with.

evaluations (but as I have already argued, one can easily exaggerate the importance of this distinction where real policy learning is concerned).

When done well, both the macro- and microvariants of this "diagnostic" approach rely on explicit theorizing. Pragmatism does not imply absence of theory. The only meaningful way to sift through the evidence—or indeed to know what kind of evidence to look for—is through the prism provided by clearly articulated theoretical frames. Where pragmatism comes in is with the analyst's willingness to shift from one model of the world to another as the evidence accumulates, and with his or her proclivity to experiment with different potential policy solutions.

Perhaps the best way to bring this micro-macro convergence into sharper relief is to describe how it differs from other ways of thinking about reform. Here is a stylized, but (hopefully) not too misleading representation of the traditional policy frame that the new approach supplants:

—The traditional approach is *presumptive*, rather than *diagnostic*. That is, it starts with strong priors about the nature of the problem and the appropriate fixes. On the macro front, both import-substituting industrialization and the Washington Consensus, despite their huge differences, are examples of this frame. On the social policy front, the U.N. Millennium Project is a good example insofar as it comes with ready-made solutions—mainly an across-the-board ramping up of expenditures on public infrastructure and human capital—even though Jeffrey Sachs would presumably argue that the project's recommendations are based on highly context-specific diagnostic work.

—It is typically operationalized via a *long list of reforms* (the proverbial "laundry list"). This is true of all the strategies mentioned in the preceding paragraph. When reforms disappoint, the typical response is to increase the items on the list rather than question whether the problem may have been with the initial list.

—It emphasizes the *complementarity* among reforms rather than their sequencing and prioritization. So trade liberalization, for example, needs to be pursued alongside tax reform, product-market deregulation, and labor-market flexibility. Investment in education has to be supported by investments in health and public infrastructure.

—There is a bias toward *universal recipes, best practices, and rules of thumb*. The tendency is to look for general recommendations and "model" institutional arrangements. Recommendations tend to be poorly contextualized.

The new policy mind-set has the following characteristics:

—It starts with *relative agnosticism* as to what works and what does not. It is explicitly *diagnostic* in its strategy to identify bottlenecks and constraints.

—It emphasizes *experimentation* as a strategy for discovery of what works. *Monitoring* and *evaluation* are essential in order to learn which experiments work and which fail.

—It tends to look for *selective, relatively narrowly targeted reforms*. Its main tained hypothesis is that lots of "slack" exists in poor countries. Simple change can make a big difference. In other words, there are lots of $100 bills on th sidewalk.

—It is suspicious of best practices or universal remedies. It searches instead fo *policy innovations* that provide a shortcut around local second-best or politica complications.

Here is a litmus test to separate adherents to these two policy frames: "Do you believe there is an unconditional and unambiguous mapping from specific *policies* to economic outcomes?" If the answer is yes with little hesitation, then the individual is in the presumptive camp. If inclined to say no, one is a fellow traveler of the experimentalists.[40]

What, then, does it mean to be a macrodevelopment economist and an experimentalist at the same time? There is no contradiction here as long as "experimentalism" is interpreted broadly and not associated solely with randomized evaluations. Experimentalism in the macro context refers simply to a predisposition to find out what works through policy innovation. The evaluation of the experiment need be only as rigorous as the policy setting allows. Some of the most significant gains in economic development in history can in fact be attributed to precisely such an approach.

What I have in mind, of course, is China's experience with experimental gradualism. As Martin Ravallion recently noted: "Anyone who doubts the potential benefits to development practitioners from evaluation should study China's experience at economic reform."[41] The type of evaluation that Ravallion is referring to is not randomized field trials.

> In 1978, the Communist Party's 11th Congress broke with its ideology-based approach to policy making, in favor of a more pragmatic approach, which Deng Xiaoping famously dubbed the process of "feeling our way across the river." At its core was the idea that public action should be based on evaluations of experiences with different policies: this is essentially what was described at the time as "the intellectual approach of seeking truth from facts." In looking for facts, a high weight was put on demonstrable success in actual policy experiments on the ground. The evidence from local experiments in alternatives to collectivized farming was eventually instrumental in persuading even the old guard of the Party's leadership that rural reforms could deliver higher food output. But the evidence had to be credible. A newly created research group did field work studying local experiments on

40. For a positive model of the choice that governments face between experimenting through policy innovation and emulating "best practices" from elsewhere, see Mukand and Rodrik (2005).

41. Ravallion (2008b).

> the de-collectivization of farming using contracts with individual farmers. This helped to convince skeptical policy makers (many still imbued in Maoist ideology) of the merits of scaling up the local initiatives. The rural reforms that were then implemented nationally helped achieve probably the most dramatic reduction in the extent of poverty the world has yet seen.[42]

Not much is said about the nature of the fieldwork undertaken, but presumably it would not have satisfied the standards of the Poverty Action Lab. Nonetheless, Ravallion is undoubtedly correct in pointing to the Chinese example as perhaps the crowning achievement of the method of experimentation combined with evaluation. Some of the experiments that proved extremely successful were the household responsibility system, dual-track pricing, township-and-village enterprises, and special economic zones. "Seeing whether something worked" is hardly as rigorous as randomized evaluations. But it would be silly to claim that Chinese policymakers did not learn something from their experiments.

The experimentalist mind-set was deeply ingrained in China's approach to reform. As Sebastian Heilmann notes, "Though ambitious central state planning, grand technocratic modernization schemes, and megaprojects have never disappeared from the Chinese policy agenda, an entrenched process of experimentation that precedes the enactment of many national policies has served as a powerful correcting mechanism."[43] Heilmann documents that Chinese-style experimentation came in three distinct forms: (1) regulations identified explicitly as experimental (that is, provisional rules for trial implementation); (2) "experimental points" (that is, model demonstrations and pilot projects in specific policy areas); and (3) "experimental zones" (specially delineated local jurisdictions with broad discretionary powers to undertake experimentation). The second and third of these are relatively better known, thanks to such important examples as special economic zones. But what is striking is that no fewer than *half* of all national regulations in China in the early to mid-1980s had explicitly experimental status (see figure 2-1).[44]

The standard policy model presumes that analysis and recommendations precede the stage of policy formulation and implementation. The experimental approach implies instead "innovating through implementation first, and drafting universal laws and regulations later."[45] Interestingly, but predictably, the share of experimental regulations has declined precipitously in the aftermath of China's joining the World Trade Organization (figure 2-1).

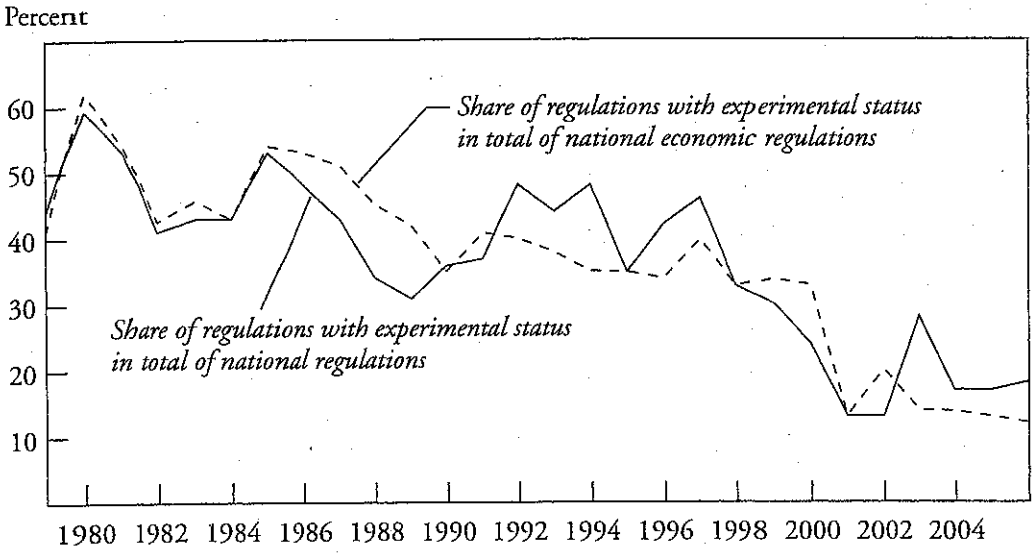
42. Ravallion (2008b, p. 2, references not included).

43. Heilmann (2008, p. 3).

44. "Experimental" in this context refers to "ordinances, stipulations, and measures issued in the name of the State Council and ministerial organs of the central government that are marked in their title as provisional, experimental or as regulating experimental points/zones." See Heilmann (2008) for further details.

45. Heilmann (2008, p. 4).

Figure 2-1. *Indicators of Policy Experimentation in China, 1979–2006*



Source: Heilman (2008).

The China example is important because it illustrates, in a vastly significant real-world instance, how the experimental approach to policy reform need not remain limited in scope and *can* extend into the domain of national policies. China is, of course, a special case in many ways. The point is not that all countries can adopt the specific type of experimentation—what Heilmann calls "experimentation under hierarchy"—that China has used to such great effect. But the mind-set exhibited in China's reform process *is* general and transferable—and it differs greatly from the mind-set behind the presumptive strategies outlined in this chapter. It illustrates perfectly the potential convergence between the ideas of microdevelopment economists and macrodevelopment economists. One would hope that the response of microexperimentalists to China's experimentalism is not to say, "But this is worthless; none of the experiments were evaluated rigorously through randomization," but to say instead, "Great, here is how economists' ideas can make the world a better place, not just one school or health district at a time."

Concluding Remarks

The practice of development economics is at the cusp of a significant opportunity—not only for a reunification of the field, long divided between macro- and microdevelopment economists, but also for a progression from presumptive approaches with ready-made universal recipes to diagnostic, contextual approaches based on experimentation and policy innovation. If carried to fruition, this transformation would represent an important advance in how development policy is carried out.

Making the most of this opportunity will require some further work. Macrodevelopment economists will have to recognize more explicitly the distinct advantages of the experimental approach, and a greater number among them will have to adopt the policy mind-set of the randomized evaluation enthusiasts. As the Chinese example illustrates, extending the experimental mind-set to the domain of economy-wide reforms is not just possible; it has already been practiced with resounding success in the most important development experience of the present generation. Microdevelopment economists, for their part, will have to recognize that one can learn from diverse types of evidence, and that while randomized evaluations are a tremendously useful addition to the empirical toolkit, the utility of the evidence they yield is restricted by the narrow and limited scope of their application. Above all, both camps have to show greater humility: macrodevelopment economists about what they already know, and microdevelopment economists about what they can learn.

References

Angrist, Joshua, and others. 2002. "Vouchers for Private Schooling in Colombia: Evidence from a Randomized Natural Experiment." *American Economic Review* 92 (December): 1535–58.

Banerjee, Abhijit V. 2005. "'New Development Economics' and the Challenge to Theory." In "New Directions in Development Economics: Theory or Empirics?" edited by Ravi Kanbur. *Economic and Political Weekly* symposium (August), typescript.

———. 2007. "Making Aid Work." In *Making Aid Work,* edited by Banerjee with others. MIT Press.

Basu, Kaushik. 2005. "The New Empirical Development Economics: Remarks on Its Philosophical Foundations." In "New Directions in Development Economics: Theory or Empirics?" edited by Ravi Kanbur. *Economic and Political Weekly* symposium (August), typescript.

Bertrand, Marianne, and others. 2007. "Obtaining a Driver's License in India: An Experimental Approach to Studying Corruption." *Quarterly Journal of Economics* 122 (November): 1639–76.

Campbell, D. T., and J. C. Stanley. 1963. *Experimental and Quasi-Experimental Designs for Research.* Chicago: Rand-McNally.

Cohen, Jessica, and Pascaline Dupas. 2007. "Free Distribution or Cost-Sharing? Evidence from a Randomized Malaria Prevention Experiment." Global Economy and Development Working Paper 11. Brookings (December).

Commission on Growth and Development. 2008. *Strategies for Sustained Growth and Inclusive Development.* Washington.

Deaton, Angus. 2007. "Evidence-Based Aid Must Not Become the Latest in a Long String of Development Fads." In *Making Aid Work,* edited by Abhijit V. Banerjee and others, pp. 60–61. MIT Press.

———. 2009. "Instruments of Development: Randomization in the Tropics, and the Search for the Elusive Keys to Economic Development" (Keynes Lecture, British Academy). Research Program in Development Studies, Princeton University (January).

Duflo, Esther. 2006. "Field Experiments in Development Economics." Prepared for the World Congress of the Econometric Society. Cambridge, Mass.: MIT Department of Economics and Abdul Latif Jameel Poverty Action Lab (January).

———. n.d. "Evaluating the Impact of Development Aid Program: The Role of Randomized Evaluations." Paper prepared for the AFD Conference, Paris, November 25.

Duflo, Esther, Rachel Glennerster, and Michael Kremer. 2006. "Using Randomization in Development Economics Research: A Toolkit" (December 12). Cambridge, Mass.: MIT Poverty Action Lab.

Hausmann, Ricardo, Lant Pritchett, and Dani Rodrik. 2005. "Growth Accelerations." *Journal of Economic Growth* 10: 303–29.

Hausmann, Ricardo, Dani Rodrik, and Charles F. Sabel. 2007. "Reconfiguring Industrial Policy: A Framework with Applications to South Africa." Harvard Kennedy School of Government (August).

Hausmann, Ricardo, Dani Rodrik, and Andrés Velasco. 2008. "Growth Diagnostics." In *The Washington Consensus Reconsidered: Towards a New Global Governance,* edited by Joseph Stiglitz and Narcis Serra. Oxford University Press.

Heilmann, Sebastian. 2008. "Policy Experimentation in China's Economic Rise." *Studies in Comparative International Development* 43 (Spring): 1–26.

Jensen, Robert, and Nolan Miller. 2008. "Giffen Behavior and Subsistence Consumption." *American Economic Review*, forthcoming.

Kanbur, Ravi, ed. 2005. "New Directions in Development Economics: Theory or Empirics?" *Economic and Political Weekly* symposium (August), typescript.

King, Gary, Robert O. Keohane, and Sidney Verba. 1994. *Designing Social Inquiry: Scientific Inference in Qualitative Research.* Princeton University Press.

Mookherjee, Dilip. 2005. "Is There Too Little Theory in Development Economics Today?" In "New Directions in Development Economics: Theory or Empirics?" edited by Ravi Kanbur. *Economic and Political Weekly* symposium (August), typescript.

Mukand, Sharun, and Dani Rodrik. 2005. "In Search of the Holy Grail: Policy Convergence, Experimentation, and Economic Performance." *American Economic Review* (March).

Over, Mead. 2008. "User Fees Can Sometimes Help the Poor." Washington: Center for Global Development (http://www.cgdev.org/doc/events/1.09.08/User_fees_can_sometimes_help_2008.pdf).

Ravallion, Martin. 2008a. "Should the Randomistas Rule?" Washington: World Bank, draft.

———. 2008b. "Evaluation in the Practice of Development." Policy Research Working Paper 4547. Washington: World Bank (March).

Rodrik, Dani. 2005. "Why We Learn Nothing from Regressing Economic Growth on Policies." Harvard University (March) (http://ksghome.harvard.edu/~drodrik/policy%20regressions.pdf).

———. 2006. "Goodbye Washington Consensus, Hello Washington Confusion?" *Journal of Economic Literature* 44 (December): 969–83.

———. 2007. *One Economics, Many Recipes: Globalization, Institutions, and Economic Growth.* Princeton University Press.

———. 2008. "Normalizing Industrial Policy." Working Paper 3. Washington: Commission on Growth and Development.

Rothwell, Peter M. 2005. "External Validity of Randomised Controlled Trials: To Whom Do the Results of This Trial Apply?" *Lancet* 365 (January): 82–93.

Sachs, Jeffrey D., Awash Teklehaimanot, and Chris Curtis 2007. "Malaria Control Calls for Mass Distribution of Insecticidal Bednets." *Lancet* 369 (June): 2143.

World Bank. 2005. *Economic Growth in the 1990s: Learning from a Decade of Reform.* Washington.