Brookings Global Economy and Development Conference
"What Works in Development? Thinking Big and Thinking Small"

THE NEW DEVELOPMENT ECONOMICS:

WE SHALL EXPERIMENT, BUT HOW SHALL WE LEARN?*

Dani Rodrik

John F. Kennedy School of Government
Harvard University

Revised Draft
May 21, 2008

ABSTRACT

Development economics is split between *macro*-development economists—who focus on economic growth, international trade, and fiscal/macro policies—and *micro*-development economists—who study microfinance, education, health, and other social programs. Recently there has been substantial convergence in the policy mindset exhibited by micro evaluation enthusiasts, on the one hand, and growth diagnosticians, on the other. At the same time, the randomized evaluation revolution has led to an accentuation of the methodological divergence between the two camps. Overcoming the split requires changes on both sides. Macro-development economists need to recognize the distinct advantages of the experimental approach and adopt the policy mindset of the randomized evaluation enthusiasts. Micro-development economists, for their part, have to recognize that the utility of randomized evaluations is restricted by the narrow and limited scope of their application. As the Chinese example illustrates, extending the experimental mindset 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 our generation.

* Paper prepared for the Brookings Development Conference, May 29-30, 2008. I am grateful to Jessica Cohen, Pascaline Dupas, and Ricardo Hausmann for comments on an earlier draft.

THE NEW DEVELOPMENT ECONOMICS:

WE SHALL EXPERIMENT, BUT HOW SHALL WE LEARN?

Dani Rodrik

I Introduction

Development economics has long been split between *macro*-development economists—who focus on economic growth, international trade, and fiscal/macro policies—and *micro*-development economists—who study microfinance, education, health, and other social programs. Even though the central question that animates both sets of economists ostensibly is how to achieve sustainable improvements in living standards in poor countries, the concerns and methods of these two camps have at times diverged so much that they seem at opposite extremes of the economics discipline.

I shall argue in this paper that there are some good reasons to be optimistic about the re-unification of the field, as these sharp distinctions are eroding in some key respects. But there are also some reasons for pessimism, related to divergence in empirical methods. This paper covers both the good and the bad news.

The good news is that there is substantial convergence in the policy mindset exhibited by micro evaluation 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 be suspicious of claims to ex ante knowledge about what works and what does not work. The answer is neither the Washington Consensus nor any specific set of initiatives in health or education. What is required instead is recognition of the

contextual nature of policy solutions. Relative ignorance calls for an approach that is explicitly *experimental*, and which 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. But I will make the case that there is indeed such a convergence, that it is a significant departure from the approaches that dominated thinking about development policy until a decade or so ago, and that it represents a significant advance over the previous generation of research.

The bad news is that there has been an accentuation of the methodological divergence between macro-development economists and micro-development economists, 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 (Banerjee 2007, Duflo 2006, Duflo, Glennerster, and Kremer 2006) has greatly enriched the micro end of the field, while creating bigger barriers between the two camps. This is not just because randomization is rarely possible with the policies—such as trade, monetary, and fiscal—that macro-development economists study. More importantly, it is because of the raising of the stakes with regard to what counts as "admissible" evidence in development. The "randomistas" (as Deaton [2007] 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). As 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" (Banerjee 2007, 12).[1] Randomized field experiments provide "hard" evidence, and by and large only such experiments do. Deprived of randomized (or natural) experiments, macro-development 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, Duflo attributes the periodic shifts in policy paradigms in development and the fact that policy debates never seem to be resolved to the weakness of the evidentiary base to date. She argues that randomization provides the way out: "All too often development policy is based on fads, and randomized evaluations could allow it to be based on evidence" Duflo (n.d., 2). Similarly, Banerjee (2007) 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. When confronted with the challenge 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 response from the randomizers is: "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."

I will argue in this paper that it is actually misleading to think of evidence from randomized evaluations as distinctly "hard" in comparison to other kinds of evidence that development economists generate and rely on. This may seem an odd claim to make in light of the apparent superiority of evidence from randomized trials. As Banerjee puts it: "The beauty of

[1] This sentence is preceded by a paragraph that recognizes the weaknesses of evidence from such "hard evidence" (including external validity and feasibility of randomization), and which 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*" (Banerjee 2007, 12).

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" (Banerjee 2007, 115-116). Case closed? Well, it depends on what the evidence is needed for and how it will be used. As economists, we might be interested in how responsive farmers are to price incentives or whether poor educational outcomes are driven in part by lack of information about relative school performance. Policy makers 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 is never assured—this is the problem of "external validity"—and it is certainly not established by the evaluation itself.[2]

Below I will discuss the issues raised by generalizability using as an illustration a recent paper by Cohen and Dupas (2007), which evaluates an experiment in Western Kenya on distribution of insecticide-treated bed nets to pregnant women. The paper finds that free distribution was vastly more effective than charging a small fee for the bed nets. As such, it represents a convincing debunking of the commonly held view that the valuation and usage of bed nets must increase with price—at least in the specific setting in which the experiment was carried out. But do the results extend to other settings in Africa as well? One can certainly make the case that it does, but the arguments one would need to deploy are perforce informal ones and they are convincing to varying degrees. In fact such arguments are not too different in kind from those that a researcher may offer in defense of a set of instrumental variables employed in a conventional econometrics study with weaker internal validity. And what is striking about the

[2] See also Basu (2005) for a very useful discussion of the limitations of randomized evaluations.

public discussion that followed the dissemination of the Cohen-Dupas paper is the wealth of reasons opponents of free distribution could offer 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 comes 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 question we need to pose of any piece of research is the Bayesian one: does the finding change our priors on the question we are interested in? Randomized evaluations do pretty well when they are targeted closely at the policy change under consideration, but less so when they require considerable extrapolation.[3] In the latter case, evidence from randomized field experiments need not be more informative than other types of evidence which 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 than others on this score than others, and deserve more of our attention on that account. 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 our priors on most of the important questions in development economics.

[3] For example, the study by Bertrand et al. (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.

But this paper is not meant to be a critique of randomized evaluations, which have indeed greatly enriched our empirical toolkit.[4] It is instead a plea for not letting prevailing methodological differences overshadow the larger convergence. My purpose is to get macro-development economists and micro-development economists to see that they have much more in common than they realize. The former are increasingly adopting the policy mindset of the latter, while the latter skate on thinner ice with their empirical work than is often thought.

The main body of the paper is in two parts. In the first part, I sketch a specific policy problem—should insecticide-treated bed nets be distributed for free or at some nominal fee?—which I use as a springboard for an extended discussion on the different types of evidence that one can bring to bear on the question, including randomized evaluations (a la Cohen and Dupas 2007).[5] The next part of the paper focuses on the convergence in policy frameworks, and discusses the main outlines of what I believe is a new paradigm in the making.

II 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.

[4] 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. See the symposium edited by Ravi Kanbur (2005).

[5] 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.

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 who are targeted by conventional public health campaigns (Sachs et al. 2007).

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 what is called "cost-sharing" (Over 2008). First, it may ensure better targeting insofar as only those who are likely to use the bed nets or those at greater at risk will want to pay. 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 healthcare 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.

It is obvious that we 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 do we gather evidence about the empirical validity of these contrasting view points? Consider three strategies.

A. Reduced-form econometrics

The fundamental question here has to do with the effectiveness of different strategies in eradicating malaria. One research approach would be to try to draw inferences about this

question by looking at the pattern of correlations across regions and over time between the type of strategy employed and the malaria outcomes on the ground. So imagine we ran the following regression for Sub-Saharan Africa:

$$Y_{it} = \alpha P_{it} + \Sigma_j \beta_j P_{it} X^j_{it} + \Sigma_j \gamma_j X^j_{it} + D_i + D_t + \varepsilon_{it} \quad (1)$$

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^j_{it} is a set of conditioning variables (income level, population density, demography, other health indicators, etc.), 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 to be discussed below, this regression can tell us how effective difference 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 $\hat{\alpha}\,(P' - P) + \Sigma_j \hat{\beta}_j\, X^j_{it}\,(P' - P)$, where $\hat{\alpha}$ and $\hat{\beta}_j$ are the estimated parameters from the regression above.

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

$$Y_i = \alpha P_i + \Sigma_j \beta_j P_i X^j_i + \Sigma_j \gamma_j X^j_i + \varepsilon_t \quad (2)$$

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

Second, how do we code and create a quantitative index for the type of policy in place in different regions or countries? Cost-sharing strategies come in many different guises, and in any case, few programs will be of the pure free-distribution or cost-sharing types. In addition, we have to 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 importantly, any regression of this type will be open to the criticism that the right-hand side variables, and 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 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 micro-econometric 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. We can generate any pattern of correlation we want by specifying those reasons and their cross-sectional variation appropriately (Rodrik 2005).

What is likely to happen in practice following an empirical exercise of this sort is a conversation and debate among 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, we can (or should) imagine that our 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 we would *never* learn anything from a regression of this kind, regardless of the quality of the supporting argumentation.

B. Qualitative evidence: surveys

One of the potential problems with 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 we need to identify the effects we are after.

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

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

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

(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 (i.e., 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 be not statistical analysis, but taking stock of the state of "local knowledge"—what people closest to the problem think—on the key questions that determine the relative effectiveness of free distribution versus cost-sharing. And open-ended questions such as (4) 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 King et al. (1994) have argued, good qualitative studies use the same logic of inference as quantitative ones. In this particular instance, we need to understand that interviewees have limited knowledge, that they have their own preconceptions (which may or may not be idiosyncratic), that they have a stake in the outcome (which may affect the nature of their responses), and that the environment in which they operate will shape their views. But even with these limitations, we ought to be able to learn something from the responses we get. 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 can not be dismissed altogether.

Of course, conclusions from such research would naturally be contested. How representative were the interviewees and can we really expect them to predict accurately the consequences of this or that program? But the relevant question here is not whether the interviews can give us a definitive answer; it is whether they can move our 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.

C. Randomized field evaluation

Finally, consider undertaking a field experiment where we randomize across recipients on whether they get ITNs for free or at a (subsidized) price. This way we can look directly for any differential effects in uptake and usage. That is exactly what is done in recent research by Cohen and Dupas (2007). These authors worked with 20 prenatal clinics in Western Kenya to offer ITNs at varying prices. The clinics were randomly divided into five groups of 4 clinics, 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 checked for usage (whether the nets were hanging on beds or not) through spot checks. 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.

When I first read this study, my initial reaction was that it settled the question once and for all. Free distribution is the way to go.[6] However, further reflection and reading on the topic[7] made clear that I had overreached. 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,[8] the setting for this study was special in a number of respects:

1. 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.
2. 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 information about

[6] 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.

[7] Stimulated in part by comments on my blog post, mentioned in the previous note.

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

malaria risks. The mass-distribution argument of Sachs, by contrast, is based on free distribution to the population at large.

3. The experiment took care of supplying ITNs to the clinics, therefore isolating the supply side from the demand side of the problem. Therefore, the experiment did not test the social marketers' claim that some degree of cost sharing is important to establish sustainable supply channels at the retail level.

4. 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 except them to hold in other setting 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 $\alpha + \Sigma_j \beta_j X^j_{it}$ which also depends on the background conditions X^j_{it}. It identifies, quite accurately, the effect of policy P under one realization of X^j_{it}, but gives us no way of parsing the manner in which those background conditions have affected the outcome, and therefore does not allow us to extrapolate to other settings. That is why it is fair game to question the generalizability of the results.[9]

[9] Deaton (2007, 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 micro-econometrics: "A well executed paper goes into a particular phenomenon in a particular location in considerable depth, data permitting.

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.[10] 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 which motivated it. It is no different in that respect than the other two research strategies I have discussed previously. Despite the clean identification provided by the randomized field experiment, those who believe we 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—although the effort will now be 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 sink in. 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 our priors. But no more than that.

D. Discussion

I have hardly scratched the surface in terms of possible research strategies. One can add various other regression-based approaches such as structural econometrics or regression-continuity. One can also think of additional qualitative strategies, such as the structured case-

---

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."

[10] 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.

study approach. The point of my discussion is not to be exhaustive but to illustrate how different styles have different strengths and weaknesses. Cross-section and panel regressions have the advantage that they can have broad coverage and they can control for at least some of the background conditions explicitly. Interviews and other qualitative approaches have the advantage that they can be carried out in a more open-ended manner, allowing unanticipated new information to play a role. Randomized evaluations have the advantage that they can nail down identification within the confines of the experiment.

In the technical jargon, the research strategies I have described above 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 I described above are weaker on internal validity—but conditional on credible identification, they have fewer problems of external validity. (In the malaria illustration above, they cover all or most of Africa as a whole and they may also have a temporal dimension.)

Some advocates of randomized evaluations would argue that internal validity trumps all else. According to this perspective, there is no point in worrying about generalizability until a causal relationship is demonstrated clearly at least once (see Cook and Campbell 1963 for the canonical statement of this position in social psychology). 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, we should proceed lexicographically: conduct randomized field experiments, and fret about external validity later.

But does this make sense from a decision-theoretic standpoint? Suppose you are a policy maker who needs to figure out which strategy to adopt—*now*. Or you are a journal editor who has to decide whether a piece of research is sufficiently well done and interesting to merit publication. In both cases, the relevant question you need to evaluate is whether the research before you *changes your priors on the question of interest*. This requires you to apply the internal and external validity tests simultaneously. Identification alone is not enough, unless you are given 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, you are 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. You are interested in knowing 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*.[11]

[11] 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

This is also in line, I think, with the revealed preference of the economics profession, which is to think of identification in terms of 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. The implication is that we cannot rank order the information content of these different kinds of studies on an a priori basis. The weights that we should 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 non-randomized 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). We 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.

Practitioners of randomized field evaluations do recognize of course problems of external validity. Duflo et al. (2006) in particular provide an excellent and comprehensive discussion of external validity pitfalls in randomized trials. As Duflo (n.d., 27) 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." 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

---

certainly not arguing that "the high quality evidence" from the randomized evaluation should be treated as "irrelevant."

"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 (2007) essay on *Making Aid Work*. Banerjee here takes the World Bank to task for producing a sourcebook on empowerment and poverty reduction in which only one of the recommendations is based on a randomized trial (school vouchers, which has been subjected to randomized evaluation in Colombia). He criticizes the recommendation on legal reform, for example, because 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" (Banerjee 2007, 14). He faults the Bank both for not showing more enthusiasm for programs like vouchers (for which we have a study with good internal validity) and for endorsing strategies like legal reform (for which we have 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" (Banerjee 2007, 13, emphasis added).

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 we were to alter, as we necessarily have to, the target population (children of secondary-school age in Colombia's low-income neighborhoods)? Or the environment in which the experiment was conducted (e.g., the availability and quality of nearby private educational facilities)? Or some details of the program (e.g., the share of private-school costs covered by the voucher)?[12] We do not know. So

[12] The study in question is Angrist et al. (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" (1556).

it is not at all clear that our priors on the relevant policy question—what strategies are worth pursuing to empower the poor and reduce poverty across the globe?—should be moved more by the Colombia study than by the cross-national studies on legal institutions. The right way to present this would have been to recognize that evidence of both types has strengths and weaknesses when it comes to informing policy makers 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 you were 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 non-experimental—paper 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 paper. Robustness issues take a whole separate section. Possible objections are anticipated, and counter-arguments 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 which may have influenced the outcomes and which do or do not exist elsewhere, they are offered in passing and are not central to the flow of the argument. Most importantly, 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.[13] And since the title, summary, motivation and conclusions of the study typically revolve around the *general* policy question, the careless reader may well walk away from the study thinking that she has learned more about the broader policy question of interest than she 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 here 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" (Rothwell 2005, 82). The long list of evidence adduced in support of this argument makes for interesting reading, in light of the parallels with current practice in economics, and I reproduce it here in the accompanying box. 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 enough times so that we feel confident in drawing general lessons. Repetition would surely help. But it is not clear that it is 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 doesn't. Absent a full theory that is being put to a test, it is somewhat arbitrary to determine under what different

[13] The version of the Cohen-Dupas (2007) paper that is online as of this writing (May 19, 2008) (http://www.brookings.edu/~/media/Files/rc/papers/2007/12_malaria_cohen/12_malaria_cohen.pdf) contains stronger language in its introduction and conclusions warning against extrapolation to other settings.

conditions the experiment ought to be repeated. If we do not have a theory of which X^j_{it}'s matter, we cannot know how to vary the background conditions. As Ravallion (2008) 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 policy options—is far from clear. The scale of the randomized trials needed to test even one large national program could well be prohibitive. (Ravallion 2008, 19)

Box: 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 US 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 UK Medical Research Council, makes virtually no mention of issues related to external validity.
- Guidance from ethics committees, such as that from the UK 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).

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 NGOs 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 (Rothwell 2005).

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, I think, 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.[14]

[14] 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 which theory suggested is most conducive to locating it (very poor Chinese consumers facing variation in the price of their and staple foods, rice or noodles). As a byproduct, the analysis clarifies the circumstances under which their result would generalize.

III The good news: convergence in policy mindsets

Abhijit Banerjee writes: "what is probably the best argument for the experimental approach [is that] it spurs innovation by making it easy to see what works" (Banerjee 2007, 122). The premise 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 the much thinking in the area of growth. Up until a decade or so ago, macro-development economists thought they had a fairly good idea about 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 clearcut 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, macro-development economists operate today in a very different intellectual environment. Gone is the confidence that we have the correct recipe, or that privatization, stabilization, and liberalization can be implemented in similar ways in different parts of the world (see World Bank 2005 and Rodrik 2006). 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 (World Bank 2005) 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'…." (World Bank 2005, xiii).

My own work (with colleagues Ricardo Hausmann, Lant Pritchett, Charles Sabel, and Andres Velasco) has focused on developing methodologies for designing country-specific growth strategies (Rodrik 2007, Hausmann et al. 2005, 2008) and on innovations in institutional arrangements for industrial policy (Rodrik 2008, Hausmann et al. 2007). We formulate the underlying problem as one of "growth diagnostics": how do we discover the binding constraints on economic growth in a specific setting, and then how do we 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 we would expect to observe under those hypotheses. Policy design in turn relies less on "best practices" and more on experimentation and monitoring.

These ideas may have been new in the growth context, but in fact they run parallel to the thinking that is reflected in the work of micro-development economists focusing on randomized evaluations. For me the epiphany occurred during an executive program we were offering at the Harvard Kennedy School on "New Thinking on Economic Growth and Development." I was sitting in a discussion that Abhijit Banerjee was conducting on the health crisis in Rajasthan and possible responses to it (which had been preceded by an excellent video produced by Banerjee and his colleagues). Over the course of the discussion, it became clear that the approach Banerjee was taking the class through was virtually identical to the Hausmann-Rodrik-Velasco "diagnostic" approach—albeit in a very different setting. You start with no presumption that that you have the answer (that poor health outcomes are due to inadequate public spending, say, or

ignorance about the value of health). So you do surveys, interviews, and collect information. You develop stories about what may account for the troubles: Are people not receiving good health care because there are no health clinics nearby? Because they do not think clinics are useful? Because there are "crack" doctors that provide apparently substitute services? Or because nurses and doctors are frequently absent? Each one of these stories has implications for the patterns you should see in your surveys and the response people give in the interviews (they throw out different "diagnostic signals" in HRV terminology). If poor people spend a considerable share of their budget on health, for example, it is unlikely that do not value it sufficiently. This kind of analysis helps you narrow the list down to a smaller list of real problems ("binding constraints"). Then you get creative and try to come up with ways—often quite unconventional—in which you can overcome these problems (lentils in exchange for inoculation, cameras in the classroom, and so on). Finally, you subject these ideas to rigorous evaluations through randomized experiments and amend them as required.

This thought process captures fairly well the spirit in which growth diagnostics exercises are supposed to be carried out as well. What my colleagues and I had begun to advocate for macro-development economists was exactly the same kind of open-minded, open-ended, pragmatic, experimental, and contextual approach. If our ideas seemed (at the time, but perhaps no longer) unorthodox, it was largely because there was already a Washington Consensus to contend with. By contrast, the absence of an equally well-formed consensus for social policy left greater space for experimentalist approaches in that domain. The main difference, of course, is that our policy innovations cannot be subject to randomized evaluations (but as I have already argued, one can easily exaggerate the importance of this distinction where real policy learning is concerned).

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 overly misleading representation of the traditional policy frame which 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 in the form of a *long list of reforms* (the proverbial "laundry list"). This is true of all the strategies mentioned in the previous item. 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.
- It exhibits a bias towards *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 mindset by contrast has the following characteristics:

- It starts with *relative agnosticism* on what works and what doesn't. 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 maintained hypothesis is there exists lots of "slack" in poor countries. Simple changes can make a big difference. In other words, there are lots of $100 bills on the sidewalk.
- It is suspicious of "best-practices" or universal remedies. It searches instead for *policy innovations* that provide a shortcut around local second-best or political 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 you answer "yes" with little hesitation, then you are in the presumptive camp. If you are inclined to say "no," then you are a fellow traveler of the experimentalists.[15]

But what does it really mean to be a macro-development economist and an experimentalist at the same time? There is no contradiction here as long as we interpret "experimentalism" broadly, and we do not associate the term 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.

[15] See Mukand and Rodrik (2005) for a positive model of the choice that governments face between experimenting through policy innovation and emulating "best practices" from elsewhere.

What I have in mind of course is China's experience with experimental gradualism. Martin Ravallion's (2008) recent paper on "Evaluation in the Practice of Development" opens with the following sentence: "Anyone who doubts the potential benefits to development practitioners from evaluation should study China's experience at economic reform." 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 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. (Ravallion 2008, 2; references not included)

We are not told much about the nature of the field work undertaken, but it is safe to presume 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 policy makers did not learn something from their experiments.

The experimentalist mindset was deeply ingrained in China's approach to reform. As Heilmann (2008, 3) notes, "[t]hough 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." Heilmann documents that Chinese-style experimentation came in three distinct forms: (1) regulations identified explicitly as experimental (i.e., provisional rules for trial implementation); (2) "experimental points" (i.e., 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 1).[16]

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" (Heilmann 2008, 4). Interestingly, but predictably, the share of experimental regulations has come down precipitously in the aftermath of China's joining the World Trade Organization (Figure 1).

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, of course, is 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.

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

But the mindset exhibited in China's reform process *is* general and transferable—and it differs greatly from the mindset behind the presumptive strategies outlined above. It is perfectly illustrative of the potential convergence between the ideas of micro-development economists and macro-development economists. One would hope that the response of micro experimentalists 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 our ideas can make the world a better place not just one school or health district at a time."

IV Concluding remarks

The practice of development economics is at the cusp of a significant opportunity. We have the prospect not only of a re-unification of the field, long divided between macro- and micro-development economists, but also of 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. Macro-development 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 mindset of the randomized evaluation enthusiasts. As the Chinese example illustrates, extending the experimental mindset 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 our generation. Micro-development 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.

In the end, both camps have to show greater humility: macro-development economists about what they already know, and micro-development economists about what they can learn.

REFERENCES

Angrist, J., E. Bettinger, E. Bloom, E. King, and M. Kremer, "Vouchers for Private Schooling in Colombia: Evidence from a Randomized Natural Experiment," American Economic Review, December 2002, 1535-1558.

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

Banerjee, Abhijit V., and others, Making Aid Work, MIT Press, 2007.

Basu, Kaushik, "The New Empirical Development Economics: Remarks on its Philosophical Foundations," in Ravi Kanbur, ed., "New Directions in Development Economics: Theory or Empirics?" a symposium in *Economic and Political Weekly*, typescript, August 2005.

Bertrand, Marianne, Simeon Djankov, Rema Hanna, and Sendhil Mullainathan, "Obtaining a Driver's License in India: An Experimental Approach to Studying Corruption," Quarterly Journal of Economics, Vol. 122, No. 4 November 2007, 1639–1676.

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

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

Deaton, Angus, 2006, "Evidence-Based Aid Must not Become the Latest in a Long String of Development Fads," in A.V. Banerjee and others, Making Aid Work, MIT Press, 2007, 60-61.

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

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

Duflo, Esther, Rachel Glennerster, and Michael Kremer, "Using Randomization in Development Economics Research: A Toolkit," December 12, 2006.

Hausmann, Ricardo, Lant Pritchett, and Dani Rodrik, "Growth Accelerations," Journal of Economic Growth, 10, 2005, 303-329.

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

Hausmann, Ricardo, Dani Rodrik, and Andres Velasco "Growth Diagnostics," in J. Stiglitz and N. Serra, eds., The Washington Consensus Reconsidered: Towards a New Global Governance, Oxford University Press, New York, 2008.

Heilmann, Sebastian, "Policy Experimentation in China's Economic Rise," Studies in Comparative International Development, Vol. 43, Issue 1, Spring 2008, pp. 1-26.

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

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

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

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

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

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

Ravallion, Martin, "Evaluation in the Practice of Development," Policy Research Working Paper 4547, World Bank, March 2008.

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

Rodrik, Dani, "Goodbye Washington Consensus, Hello Washington Confusion?" Journal of Economic Literature, XLIV, December 2006, 969-983.

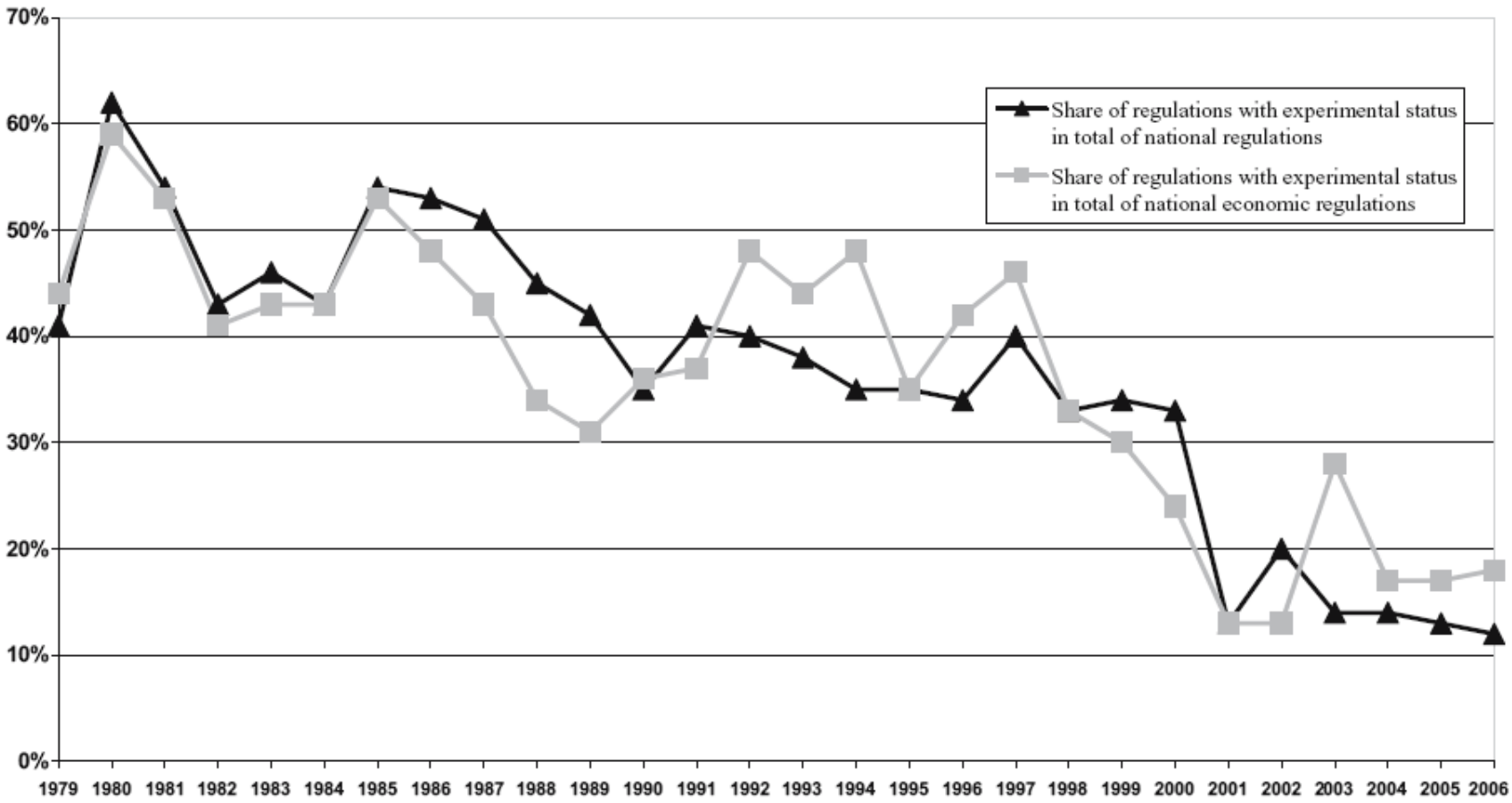
Rodrik, Dani, One Economics, Many Recipes: Globalization, Institutions, and Economic Growth, Princeton University Press, 2007.

Rodrik, Dani, "Normalizing Industrial Policy," Commission on Growth and Development, Working Paper No. 3, Washington, DC, 2008.

Rothwell, Peter M., "External validity of randomised controlled trials:
"To whom do the results of this trial apply?" The Lancet, vol. 365, 2005, 82-93.

Sachs, Jeffrey D., Awash Teklehaimanot, and Chris Curtis, "Malaria Control Calls for Mass Distribution of Insecticidal Bednets," The Lancet, June 21, 2007.

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



Source: Heilmann (2008)

Figure 1: Indicators of policy experimentation in China