1

Introduction: Thinking Big versus Thinking Small

JESSICA COHEN AND WILLIAM EASTERLY

The starting point for the contributions to this volume, and the conference for which they were prepared, is that there is no consensus on "what works" for growth and development. The ultimate goal of development research—a plausible demonstration of what has worked in the past and what might work in the future—remains elusive. As Martin Ravallion points out in his comment on chapter 2, we are beyond "policy rules" such as the Washington Consensus, and "thinking big" on development and growth is in crisis. The "big" triggers for economic growth have not been shown to work, either because they in fact did not work or because it was impossible to demonstrate their impact persuasively.

As a result, many in development have turned to "thinking small." For the most part—but not exclusively—the focus has shifted from macro- to micropolicy questions. This type of research commonly seeks the most effective method for delivering public goods such as education and vaccines. A growing methodology for analyzing micropolicy questions is randomized controlled trials (also known as Randomized Evaluations, REs). Much of this volume is about the merits and drawbacks of REs in elucidating what works in development. The specific arguments—we say more on them later in this chapter—revolve around several nagging questions. What kind of development policy research yields "hard" evidence? Are some types of evidence "harder" than others? Is there a trade-off between the scope of the questions researchers ask and the quality of the evidence

they generate? What questions and what quality of evidence matter most for development policy and aid effectiveness? In exploring these issues, it is essential to first ask how the crisis in thinking big transpired and whether thinking small is indeed a solution.

The Collapse of "Thinking Big"

The failure of thinking big—elsewhere described as "the panaceas that failed"—has been widely acknowledged.[1] Many would agree with Arnold Harberger that "there aren't too many policies that we can say with certainty . . . affect growth."[2] Some, like those behind the Barcelona Development Agenda, would go even further: "There is no single set of policies that can be guaranteed to ignite sustained growth."[3] Even the universally revered dean of growth theory, Robert Solow, believes that "in real life it is very hard to move the permanent growth rate; and when it happens . . . the source can be a bit mysterious even after the fact."[4]

Where did this pessimism come from? As both Abhijit Banerjee, and William Easterly in his comment on Banerjeee, discuss later in the volume, several contributing factors readily come to mind: despite concerted attempts, macroeconomists were unable to deliver higher growth; the credibility of the growth regression literature waned; extremely volatile growth rates could not be explained; and growth analysis neglected to do enough long-run regressions.

The Failure of Big Pushes to Raise Growth

Three unsuccessful pushes are particularly notable:

1. The early big push in foreign aid (especially in the most aid-intensive continent, Africa).

2. Structural adjustment (also known as the Washington Consensus) in the 1980s and 1990s.

3. "Shock therapy" in the former Communist countries.

All of these episodes are far from natural experiments, of course, with adverse selection posing a severe problem for the interpretation of policy impact. However, all three had such poor outcomes that the counterfactual—that growth would have been even worse without the macroeconomic intervention—was hardly plausible.

1. Easterly (2001).

2. Harberger (2003).

3. The Barcelona Development Agenda was the consensus document resulting from Forum Barcelona 2004, Barcelona, Spain, September 24–25 (www.barcelona2004.org/esp/banco_del_conocimiento/docs/CO_47_EN.pdf). Those involved in this exercise included Olivier Blanchard, Guillermo Calvo, Stanley Fischer, Jeffrey Frankel, Paul Krugman, Dani Rodrik, Jeffrey D. Sachs, and Joseph E. Stiglitz.

4. Solow (2007).

The Failure of the Growth Regression Literature

The pessimism surrounding big pushes intensified as the credibility of the cross-country growth literature declined, with its endless claims for some new "key to growth" (regularly found to be "significant") and probably well-deserved reputation for rampant data mining. As the Easterly comment on Banerjee notes, the number of variables claimed to be significant right-hand-side (RHS) determinants approached 145, which is probably an undercount.[5] Having a long list of possible controls to play with, researchers found it easy enough to arrive at significant results, and using the abundant heuristic biases that make it possible to see patterns in randomness, convinced themselves that the significant results were from the "right" specification, and that the others (usually unreported) were from the "wrong" ones.[6]

The growth literature was also criticized for its inability to address causality. In the absence of clear evidence that growth outcomes can be attributed to specific levers, development research has severely limited utility for policy. This deficiency was probably due to the infeasibility of instrumenting for multiple RHS variables. Any such attempts usually relied on the Arellano-Bond or Arellano-Bover dynamic panel techniques, which (essentially using lagged RHS variables as instruments) became a kind of magical machine churning out causal econometric results. Unfortunately, the identifying assumptions were so implausible as to leave most outside observers unconvinced. This left the causality question unresolved.

Not to overstate the inevitability of the collapse of growth knowledge, neither data mining nor causality was a completely hopeless cause in aggregate regressions. Data mining can be held in check with a well-known methodology: estimating slight variants of the original specification that are as plausible as the original; or, better yet, adding new data that were unavailable at the time of the original estimation to the exact specification. As far as causality was concerned, occasionally there would be a reasonably plausible instrument for a RHS variable of particular interest.

Both remedies can be explored in the hotly debated literature on the effect of aid on growth. Since it is hard for researchers to hold themselves aloof from the strong vested interests in and political biases for or against aid, there was enormous scope for data mining in aid and growth regressions. One very simple test of data mining is to add new data that were not available at the time of the original regression specification and see if the results still hold. The famous result of Craig Burnside and David Dollar that "aid raises growth in a good policy environment" did not pass this test.[7]

5. Durlauf, Johnson, and Temple (2005).
6. Kahneman and Tversky (2000); Kahneman and others (1982); Gilovich and others (2002).
7. See Burnside and Dollar (2000); Easterly, Levine, and Roodman (2004).

As for causality, one promising instrument for aid was population size, because of the quirk that the aid donor bureaucracy does not fully increase aid dollars one for one with recipient population size. Log of population is thus an excellent predictor of aid/gross domestic product (GDP), thankfully unrelated to the economic motivations for aid, and has been used in many studies. Another original strategy for identifying the impact of aid on growth has been to instrument aid from the Organization of Petroleum Exporting Countries (OPEC) to their poor Muslim allies with the interaction between oil price and a Muslim dummy variable.[8] This approach uncovered a short-term effect of aid on output, but also a zero effect on medium-term growth. The generalizability problems of such identification strategies are much like those of REs. Does small-population-induced aid have the same effect as other aid? Does intra-Muslim aid have the same effect as aid from the United Kingdom to Africa? Another problem, which REs typically do not have, is serious doubt about the excludability of the instrument in cross-country growth regressions.

But even when establishing causality was not possible, it would have been equally extreme to say that strong partial correlations would or should have no effect on priors about causal policy effects. Researchers were also probably influenced by the established body of theory, which tended to predict causal effects of policies on growth without much reason to think that growth would feed back into policies.

The Volatility of Growth Rates

By and large, the growth literature has also failed to establish even robust partial correlations between growth and country characteristics. One reason is that growth rates are extremely volatile while country characteristics are persistent. A crude way of showing growth volatility is to do a random effects regression on a panel of annual per capita growth rates between 1960 and 2005. This reveals that only 8 percent of the cross-time, cross-country variation in growth is due to permanent country effects; the other 92 percent is transitory. The annual standard error of the pure time-varying component of growth is an amazing 5.06 percentage points.[9] In the latest successive decades, 1985–95 and 1995–2005, there is virtually zero correlation between a country's performance in one decade and its performance in the next (hence virtually 100 percent mean reversion). This is very bad news when almost all of the plausible determinants of growth are relatively permanent country factors. It was like trying to explain differences in this week's batting averages of baseball players by differences in long-run fundamentals such as training regimens (or steroids!). The noise-to-signal ratio is so high with both

8. Werker, Ahmed, and Cohen (2009).

9. This was documented more than fifteen years ago by Easterly and others (1993), and in another form by Pritchett, Rodrik, and Hausmann (2005), who showed high-growth episodes to be almost always temporary.

weekly batting averages and decade growth rates that any such attempt is largely futile.

Long-Run versus Short-Run Development and Growth Literature

When asked about the impact of the French Revolution, Zhou En-Lai reportedly said it was "too soon to tell." So with growth performance. A *very* long-run average of growth rates is needed to lower the noise-to-signal ratio enough to have something interesting to relate. To put it another way, growth analysis has suffered from what Daniel Kahneman and Amos Tversky sarcastically call the Law of Small Numbers problem: reading too much into growth differences over one or two decades when they were mostly transitory.[10] The obvious answer was to go to more long-run analysis, which is in fact the direction the macro literature took, doing regressions for log *levels* of per capita income as a function of long-run characteristics such as institutions.[11]

The lack of persistence of growth rates made data mining easy—and easy to catch. As new growth observations came in, almost uncorrelated with past observations, the data miners would keep finding new variables that "explained" growth. As yet more new data came along, country factors would be the same, but the growth rates would get scrambled again, and the results would vanish. In an amusing nonacademic example, countries in which people ate fast were found to grow more rapidly over 2001–08 than countries in which people ate slowly.[12] Fast-Food USA has been growing faster than slow-eating Japan. The culture on eating presumably is persistent, so this result would not have held in the old days of rapid Japanese and slow U.S. growth, and any slow-eating French boom would make the result disappear again.

Those who reject such chicanery currently hold the field in empirical growth research. This position itself may not be sustainable, however. For the consumers of academic research, just saying "it's too soon to tell" about a matter as visible as economic growth differences is almost impossible to accept. So the nearly universal debunking of growth knowledge has not stopped attempts to explain growth.

The latest attempts appear to embrace a theory on the order of

$$\textit{Growth (country i, period t)} = \textit{Coefficient (country i, period t)} * \textit{Policy (country i, period t).}$$

This equation finally fits the data very well! Alas, tautological and nonfalsifiable theories are not usually allowed. Lest we appear to exaggerate, consider a statement by the World Bank Growth Commission in the aftermath of $4 million

10. Kahneman and Tversky (1982).

11. Acemoglu and others (2001, 2004); Easterly and Levine (2003); Rodrik, Subramanian, and Trebbi (2004).

12. See Floyd Norris, "Eat Quickly, for the Economy's Sake," *New York Times*, May 8, 2009.

worth of conferences and consultants: "It is hard to know how the economy will respond to a policy, and the right answer in the present moment may not apply in the future."[13] In chapter 2 of this volume, Dani Rodrik also seems to flirt with this extreme at times, although he avoids nonfalsifiable tautology by laying claim to independent knowledge with growth diagnostics exercises that would give insights into the coefficient (i,t)'s.[14]

Ricardo Hausmann in chapter 6 notes the complexity of public policy and institutions and suggests that the search for which policy is "the answer" only makes sense if there was a "central planner" implementing policies, which is no more feasible for public policy than for private goods markets. He argues for decentralized public policymaking to address such complexity.

The unpopular but well-justified focus on the long term can also generate insights into development from long-run stylized facts. A long tradition in development is to establish robust stylized facts in levels that guide development thinking—examples are the positive correlation between democracy and per capita income (which has stimulated thinking about causal channels in both directions), the negative correlation between per capita income and fertility (often interpreted causally as "development is the best contraceptive"), the relationship between life expectancy and income, and how that relationship (the "Preston curve") has shifted over time (suggesting that major changes in health technology can improve health without income growth), and so on.[15] The neoclassical production function model of development has come into question because of its violation of many stylized facts about development (such as capital flows, brain drain, and the failure of absolute convergence).[16] New growth models have also been guided by macro stylized facts. For example, idea models that stress R&D efforts as a determinant of growth have run afoul of the stationarity of growth rates and the nonstationarity of R&D efforts,[17] although they would fit the very long run of world development as a whole.[18] Another stylized fact in many analyses is the surprising significance of very long-run history for determining today's outcomes, which may lend support to some growth theory models with increasing returns and sensitivity to initial conditions. As David Weil notes, commenting on chap-

13. World Bank Growth Commission on Development (2008).

14. Of course, parameter heterogeneity is indeed a problem, and econometrics can at least discuss what the estimated coefficients mean when there is such heterogeneity. The coefficients will mean something very much like what the analogous situation in micro experiments means, with the estimate signifying something like a "local average treatment effect" (Deaton [2009]). The difference in macro regressions is that the "local" is averaging over a lot of very varied experiences across countries, while the "local" in micro is averaging over a specific population within a small treatment site.

15. On life expectancy and income, see Deaton (2006).

16. As discussed in Klenow and Rodríguez-Clare (1997); Easterly and Levine (2001).

17. Jones (2005).

18. Kremer (1993); Galor and Weil (2000).

ter 4, scientific experiments are not the only means of learning about development; historians do not do experiments, but most economists think that they learn something from historians (including economic historians such as Stanley Engerman and Kenneth Sokoloff analyzing why North America is richer than South).

So thinking big is not dead. However, sixty-year-old hopes that thinking big would translate into clear guidance on how to move immediately into rapid growth and development have been repeatedly disappointed.

The New Promise of "Thinking Small"

The macro literature is not alone in lacking decisive evidence. REs, "natural experiments," and other methodologies prioritizing transparency and clean identification became popular because of the great vacuum of microevidence on development projects. As Lant Pritchett has eloquently put it, nearly all World Bank discussion of policies or project design had the character of "ignorant armies clashing by night."[19] Despite the heated debate among advocates of various activities, they rarely presented any firm evidence or considered the likely impact of those actions. As far as we know, there was never any definitive evidence that would inform decisions of funding one instrument versus another (such as vaccinations versus public education about hygiene to improve health, or textbook reform versus teacher training to improve educational quality).

Even a World Bank handbook was quick to note that "despite the billions of dollars spent on development assistance each year, there is still very little known about the actual impact of projects on the poor."[20] The RE literature made a clear case for basing aid policy on evidence rather than prejudice and special interests. This methodology holds tremendous promise for improving aid effectiveness (and cost-effectiveness) by helping policymakers, donors, and nongovernmental organizations (NGOs) choose between a nearly infinite range of development program possibilities. While REs have some drawbacks—and doing them well is often an art—they have the undeniable strength of transparency and usability. A simple comparison of means between treatment and control groups can persuasively illustrate the impact of many different types of development policies and NGO programs.

But has the RE literature managed to solve the problems that afflicted the thinking big literature? There is already some backlash against REs, in some ways reminiscent of the backlash against aggregate macrowork, although many of the issues are quite different.

19. Pritchett (2009, p. 121).
20. Baker (2000).

Arguments for and against REs

The view that REs are a major advance over cross-country empirics has drawn strong support as well as criticism. We do not try to resolve that debate here. Instead we simply summarize some of the arguments on both sides regarding RE identification, external validity, RE links to theory, data mining, RE effect on implementing agencies, its effect on policy, and the underlying ethical and social engineering concerns. The order in which we present them should *not* be taken to imply that one side or the other has a decisive last word on the matter.

RE Identification. The most important claim of RE supporters is that they have solved the identification problem. Controlling the assignment mechanism through randomization allows a causal ("treatment") effect (usually of a development program or policy) to be estimated by removing selection bias. If the great majority of people offered the treatment in an RE accept it or (in the less likely case) there is no selective compliance with the randomization, REs allow one to estimate average program impact through a simple comparison of means across treatment and control.

When compliance with the program or policy is selective (for example, when only the sickest people take up a health product or the smartest children use new textbooks in a school), a common approach is to instrument the actual treatment with the treatment assignment. This is an advance over using standard instrumental variables (IVs) since (as long as the randomization was not compromised somehow) one can be sure that the instrument is exogenous and that the causal effect on outcomes is being identified. In sum, REs deliver an internally valid estimate of the causal effect of a policy or program on outcomes, something that is unattainable with observational studies.

A primary criticism with RE identification (when the treatment effect is instrumented), however, is the ambiguity about *what* is being identified. If the impact of the policy or program varies across members of the treatment group (that is, there are heterogeneous treatment effects), the parameter being identified is a Local Average Treatment Effect (LATE). Specifically, it is the impact for the subset of people who were induced to adopt the treatment because it was offered to them ("compliers") and excludes those who would never adopt the treatment, or who would have adopted it in the absence of the intervention.[21] Although this contains much useful information, it omits important details about the effect of the treatment on individuals who are not "local." As Ravallion points out later in the volume, this poses a problem for generalizability. If the program has a very different impact on the "compliers" than on others in the population, how can a policymaker determine its possible impact on a national level? Further, the average estimate that LATE delivers has limited usefulness. How valuable is LATE rel-

21. Imbens and Angrist (1994).

ative to the median effect of a program? And is LATE useful for the case in which a program has a positive average impact but causes a small share of people to suffer very negative consequences?

External Validity. Can an RE finding be generalized to other settings? Among the skeptics, Angus Deaton and Dani Rodrik might concede RE's internal validity but would seriously question its external validity.[22] Like many others, they wonder whether the particular outcome of a particular program carried out on a particular population in a particular country by a particular implementing organization would be found in other circumstances. One of the main objectives of REs is to guide policy decisions, but how can policymakers be sure whether a program that worked well in one country would have similar effects in another? Or, if it worked with an urban population, that it would work with a rural one? Equally important, how can one tell that the program would be as effective if implemented by the government rather than by an NGO? According to Nancy Cartwright, REs do "not tell us what the overall outcome on the effect in question would be from introducing the treatment in some particular way in an uncontrolled situation, even if we consider introducing it only in the very population sampled. For that we need a causal model."[23]

One possible solution is to replicate the program in many different settings to confirm a general result.[24] However, the incentives for researchers to do replications fall off very rapidly with the number of replications already performed. Moreover, it is unclear how many are needed or how to choose the right sample of environments (with what factors varying?) to validate a result from the original study. Replication is often mentioned as a solution to the external validity problem without guidance on *how many* or *what kind* of replications are "enough" to establish generalizability. The problems in generalizing from a small slice of experience are analogous to those for the Law of Small Numbers in aggregate growth analysis mentioned earlier.

For some, the biggest problem is the lack of a model to clarify why, when, and where the treatment is expected to work.[25] In other words, an RE is most useful when it sheds light on some behavioral response (such as the price elasticity of demand for health inputs)—although even then it may not extrapolate to other settings. REs are less useful when they issue a blanket claim that "*X* works but not *Y*" on the basis of one very small sample in a particular context, without any clear intuition as to why *X* is more likely to work than *Y*. As Rodrik points out in chapter 2, the progression from RE results to policy often involves the same kinds of appeals to theoretical priors, common sense, casual empirics about similarity of the new policy setting to the original research setting in some (but not all) aspects,

22. See Deaton (2009); and Rodrik's discussion in chapter 2.
23. Cartwright (2007) in Deaton (2008).
24. Duflo, Glennester, and Kremer (2008); Banerjee and Duflo (2008).
25. See, for example, Deaton (2009).

and other more casual sources of evidence as does using aggregate econometric results and stylized facts to influence policy.

In their defense, however, REs are not alone in facing a trade-off between internal validity and generalizability. Although cross-country macrostudies are widely thought to be more generalizable than REs because they estimate averages over time, space, and population (see chapter 2), some would counter that observational studies of program impact covering large temporal and cross-sectional dimensions are equally prone to this trade-off.[26] Such studies must often control for multiple covariates (or use matching) to estimate a treatment effect, but, once covariates are controlled for, the estimate will be dominated by groups with overlapping covariates. In other words, cross-temporal/cross-sectional studies do not estimate an average treatment effect for an entire population—but only for a certain subpopulation—just as in the case of REs and IV studies. The advantage of REs here is that they often produce detailed microdata that can help in understanding this population and assessing its generality.

Furthermore, previous theoretical and empirical research, as well as plain common sense, provides some intuitive grounds for determining which characteristics would really affect program impact (and which would be inconsequential or secondary) and hence could provide guidance for a manageable number of replications. In an evaluation of a school intervention that tries to reduce student-teacher ratios, for example, the main concern would be whether the results extend to a context in which teachers' contracts are different, not the color of the school walls. This sense of what should matter and what should not pervades today's increasingly thorough and careful research into REs and external validity. As Banerjee points out in chapter 7, the significant advantage of REs lies in replicability, since hypotheses about external validity are *testable* with this methodology. Ideally, one uses theory to decide what factors would matter for replicability, but absent that theory, REs can be repeated as frequently as necessary. Atheoretical replication is clearly not ideal from a social science perspective, but to policymakers or NGOs wanting to know how to improve education or reduce poverty in their country, this feature of REs is a great help.

Of course, it is also important to consider "confounding" factors in assessing external validity, as illustrated in a study of subsidized bed nets in Kenya.[27] The larger policy question that motivated this investigation was whether bed nets should be free or highly subsidized for pregnant women. Two factors suggested that the study's results might not hold for other populations: (1) social marketing of bed nets had taken place recently, and (2) only pregnant women were targeted with bed nets. Although these factors were likely to have influenced the behavioral response to bed-net pricing, they are not necessarily a problem for external validity. From a

26. See Imbens and Woolridge (2008); Banerjee and Duflo (2008).

27. See Cohen and Dupas (2009); Rodrik (chapter 2).

policy perspective, this is the relevant environment and population to consider: social marketing of bed nets is exceedingly common in Africa, and pregnant women and their babies are the target groups for malaria prevention.

In any case, exact replication of REs in different contexts is not required to draw broad lessons about development, as demonstrated by Michael Kremer and Alaka Holla in chapter 4, in their discussion of RE literature dealing with education and public health and with price sensitivity in these areas. They find that a number of REs actually explore program variation within different treatment arms of the same experiment.[28] Furthermore, a number of studies have used randomization to estimate structural models.[29] A better grounding of REs in economic theory could no doubt elucidate the environmental and behavioral factors behind the results being reported in the program evaluation literature.

Another possibility is to use qualitative research to explore the results of an RE more closely and to give direction for future research. Anne Case (this volume) argues that REs miss out on important information by focusing only on quantitative variables. She suggests a mix of qualitative and quantitative analysis. Of course, macrostudies would benefit from this mix as well, and one advantage of REs is that they routinely collect very detailed microdata that could be combined with a qualitative analysis. Nava Ashraf goes a step further, arguing that qualitative analysis should be combined with theory to help use REs more systematically in a search for what works in development. Responding to Hausmann's claim in chapter 6 that REs cannot possibly guide complex and multidimensional policies, Ashraf argues that qualitative evidence can shed some light on the many ways in which quantitative RE results may vary across contexts and then theory can be used to predict which types of variation would meaningfully affect the estimated program impact.

Another way forward for REs is to make better use of macrodata. Using the example of determinants of child mortality, Peter Boone and Simon Johnson argue in chapter 3 that—in the absence of a theory that is being tested—correlations observed in macrodata can be very useful for motivating the design of an RE. Klenow (commenting on chapter 7) points out that a divide between "micro" (controlled, laboratory work) and "macro" (data analysis of trends, for example) exists in most areas of the natural sciences, but that it is routine for macrostudies to motivate and guide micro ones. He uses the example of macrostudies of trends in obesity guiding lab experiments on diet and exercise, or the correlation between smoking and lung cancer spurring more controlled microwork on carcinogenic triggers.

RE Links to Theory. Should REs stop making general "X works"–type statements and instead try to estimate parameters in a theoretical model of human behavior, as critics suggest? For Deaton, "heterogeneity is not a technical problem,

28. See Banerjee and Duflo (2008).
29. See Imbens (2009).

but a symptom of something deeper, which is the failure to specify causal models of the processes we are examining."[30] The shift toward such models is already under way in the RE literature but is not reflected much in this volume. The extensive discussion of the free provision of bed nets and water purification tablets, for example, provides little theoretical analysis and instead jumps immediately to policy: bed nets should be given away free to avoid reduction in uptake; on the other hand, a fee for water purification tablets would help to screen out likely nonusers (see chapters 3 and 4 and the comments by Jessica Cohen and David Weil).

In chapter 4, Kremer and Holla do identify an interesting theoretical anomaly: poor people seem remarkably price-sensitive toward goods that should presumably pay for themselves in the form of health and lost income. They consider two explanations—(1) a behavioral anomaly (such as commitment problems or hyperbolic discounting, "procrastination"), and (2) a lack of knowledge of the payoff from these goods—and appear to favor the first. David Weil notes how problematic this conclusion is: are the same people also procrastinating on getting their crops into the ground or harvested? If not, why is this behavioral anomaly focused on health?

These alternatives have very different policy implications, since zero pricing might "nudge" the behavior in the right direction under the first hypothesis but would just mean a lot of the goods would go unused under the second. Anecdotes exist of malaria bed nets being used as wedding veils and fishing nets, for example, and even more systematic evidence reveals that free insecticide-treated bed nets donated by an NGO have been diverted for drying and catching fish on Lake Victoria in western Kenya. Given the drastically different implications of the two explanations, the debate surrounding them has attracted surprisingly little attention, as attested by the underdevelopment of theory in the RE literature. Hypothesis 2 creates even more puzzles: why do poor people not have accurate knowledge of the payoff to health goods when there is a strong incentive to acquire such knowledge and the knowledge seems readily available? Efforts to address this question could lead to even more interesting theory and empirics about knowledge acquisition. Do the poor in Africa not place much credence in scientific medicine because they have a malfunctioning health system that does not reliably deliver benefits from scientific medicine? Does more general education correlate with more knowledge of payoffs to health goods? As David Weil notes, maybe the customers do not believe the mosquito theory of malaria, doubt the quality of the bed net, or have trouble knowing whom to believe between bed-net promoters and traditional healers.

The overall result of such efforts could shed light on the long-debated issue of whether poor people fit the rational *homo economicus* model (at a time when no

30. Deaton (2009).

one is even sure that highly educated rich people do). Perhaps the paucity of such discussion reflects a fear of political incorrectness, but it may also reflect an insufficient commitment to theoretical inquiry in the RE literature.[31] As Captain Von Trapp says in the *Sound of Music,* we all seem to be "suffering from a deplorable lack of curiosity." Kremer, for one, seems ambivalent on the issue of rational decisionmaking among the poor: although he presumes an extremely high degree of rationality and calculation among poor shopkeepers, from which he reaches an estimate of the return to capital, he questions rationality in usage of health services and fertilizer.[32]

Arguably, the weak links of REs to theory—and the consequences that might have for generalizability—are relevant for observational macrostudies as well. Sendhil Mullainathan (this volume) argues that the absence of theory compromises generalizability both when one is seeking to extrapolate "up" from a microstudy and "down" from a macrostudy. He uses the example of wanting to know what average wages are in Oklahoma. Which information is preferable: average wages in Kansas or average wages in the United States overall? Neither is sufficient for extrapolating wages in Oklahoma without some theory to guide that extrapolation.

While a better grounding in theory would benefit observational studies as well as REs, the latter do have several advantages in their links to theory: REs can test theories in a very controlled environment, generate parameter estimates with much more internal validity than other approaches, and deliver a body of evidence on which a theory can be developed. It is possible to get an accurate estimate of price sensitivity to bed nets or water purification because of RE methodology. Behavioral theories consistent with that evidence can be developed precisely because the RE methodology offers replicability and transparency. These behavioral theories can in turn be tested with REs. The circle from evidence generation to theory back to evidence—facilitated by the experimental methodology—occurs frequently within experimental economics. The early economic experiments such as the ultimatum game generated a body of evidence contradicting the rational actor model, which in turn led to a proliferation of theoretical models of fairness and reciprocal behavior.[33] These alternative models generate behavioral parameters that can be

31. Several issues could be discussed on this specific "human capital irrationality" result: (1) the effect of education on the ability to process other information about human capital, such as that on nutrition and transmission of malaria and diarrhea; (2) how "human capital irrationality" is related to what seems to be high rationality and resourcefulness in managing finances (Collins and others [2009]); and (3) whether the delivery of health and nutrition services to the poor can affect their acceptance of scientific theories of disease.

32. See Kremer, Lee, and Robinson (2008); Kremer and Holla (chapter 4); and Duflo, Kremer, and Robinson (2007).

33. See, for example, Duflo, Kremer, and Robinson (2009), who use early experimental evidence on the importance of perceived "fairness" in the context of moral hazard to design a principal-agent model incorporating fairness, which is then tested experimentally. (They then use all of these experimental results to motivate a model of "inequity aversion"!)

calibrated by future experiments. Thus although the criticism that REs ought to be tied more closely to theory is a valid one, it fails to recognize that REs are arguably the best-suited methodology for linking with theory.

Finally, pragmatically, one may not always want to link REs to theory. In many areas of development, one may simply want to know whether a policy or program works or does not work. Obviously it is optimal if this evidence feeds into a theory that can inform other research, but if it is used to reallocate many millions of dollars in foreign aid, that is not such a bad outcome.

Data Mining. The pure RE design prevents data mining, for there is only one regression, of outcome on the treatment dummy, which is specified in advance. This sounds like a clear advantage for RE over growth regressions with almost infinitely flexible specifications.

However, as many have pointed out, the incentives for a "result" are still very strong, and the RE design is not quite as restrictive as just stated.[34] First, there are numerous possible outcome measures. Second, more than one site may be reporting results. Third, RE regressions usually include covariates, and the list could be almost as long as in growth regressions. If the randomization is successful, the inclusion of covariates should not change the coefficient but only increase precision. In the presence of budget and administrative constraints, however, RE samples are often small, which means it is more difficult to achieve a balance across treatment groups, and covariates can change the estimated treatment effect. Data mining is most likely to occur in the search for a significant program impact among different subpopulations. After all the expense and time of a randomized trial, not only is it is very hard to conclude "we found nothing," but with the aforementioned tendency to see patterns in randomness, it becomes difficult to resist the temptation to play with all these margins to get a result. Those who acknowledge these problems recommend full disclosure, but this is hard to enforce.[35] In chapter 3, Boone and Johnson argue that the current RE methodology used in development economics would not meet the standards of the medical literature largely because it fails to prespecify primary "endpoints" and typically lacks a statistical analysis plan.

All the same, REs are moving toward higher standards with regard to data mining. The majority of well-executed REs report the basic specification without covariates. Most REs use ex ante stratification as well, which allows them to include subgroup analyses that preserve the integrity of the randomization. Further, while the ex post search for significant program effects is not exactly kosher, it helps pave the way forward, by suggesting follow-up evaluations and hypotheses to be tested.[36] Even when REs are not done as carefully as they should be, the scope for data mining is likely to be less than in cross-country regressions.

34. See Deaton (2009); Duflo, Glennerster, and Kremer (2008).
35. See Duflo, Glennester, and Kremer (2008).
36. Imbens (2009).

Perhaps a larger problem is publication bias. If a researcher is so virtuous as to disclose a "no results" finding, the study will probably not get published (especially since such an outcome does not necessarily prove zero effect but may merely mean the estimated effect suffers from imprecision). Somebody later surveying the literature might then assume a positive and significant result of the intervention in question, unaware of the unpublished regressions with no results. It would be nice to see an aggregate test of the RE literature for data mining/publication bias, as has been done with other literatures. Of course, publication bias is a concern in all empirical literatures, not just the RE literature—in any context it can make it difficult to infer "what works" from published information. This problem could be ameliorated if institutions could conduct REs without an (academic) publication incentive.

The Effect of REs on Implementing Agencies. RE will inevitably be seen as an evaluation of, at the very least, whether *that* program by *that* NGO or aid agency (or specific department or even individual) worked on that occasion. This will in many circumstances reflect well or badly on those proposing and implementing the program. RE proponents such as Esther Duflo and Michael Kremer want to discourage this interpretation. Furthermore, they oppose any scheme that would reward or penalize particular aid actors for positive or negative results of evaluations, in part because implementing agencies might then be less likely to cooperate in an RE.[37] Or if the agency felt threatened by a negative result or perceived great rewards to a positive result, it might manipulate the results.

Even so, many RE proponents do think that aid systems could be redesigned to reward positive REs: "Positive results . . . can help build a consensus for the project, which has the potential to be extended far beyond the scale that was initially envisioned."[38] If this is true, it is hard to imagine that an implementing agency or its staff would be indifferent to a large increase in its budget from scaling up, or to kudos for having found a very successful intervention. Official aid agencies and NGOs are notoriously sensitive to good or bad press. It would be naïve to think that they are ever indifferent to how an evaluation comes out. Hence the incentive for implementing agencies to manipulate results already exists. REs in medicine (the gold standard for RE literature) have been criticized on these very grounds when private drugmakers finance RE studies of their drugs.[39] When agencies are already confident a program is working, notes Ravallion later in the volume, they selectively agree to REs, which could bias the probability of a positive evaluation upward. The RE literature obviously needs to

37. See the discussion on the Creative Capitalism website: "Holding Aid Agencies Accountable," an e-mail exchange between William Easterly, Esther Duflo, and Michael Kremer, July 31, 2008 (http://creativecapitalism.typepad.com/creative_capitalism/2008/07/exchange.html).

38. Duflo (2004).

39. Deaton (2009).

devote more attention to such manipulation of evaluations and ways to counteract this threat.

Of course, not all REs are directed at existing programs—that is, ongoing programs whose implementers may have a stake in finding positive results. Many evaluate existing government policies or new programs that are run by the researchers themselves.[40] When REs are involved with specific implementing agencies, they often deal with new variations on existing programs.[41] Although the desire for positive press may still be present in these cases, there is much less incentive to interfere with the study so as to influence outcomes. To assuage an implementing NGO's fear of negative results, an RE may evaluate several versions of a program simultaneously, shifting the focus to what works best from whether a program works at all. The testing of new variations can relieve the pressure of demonstrating that sunk resources were spent wisely.

If the anticipation of a negative evaluation of an existing program could influence the validity of an RE, it would be sensible for REs to test programs or policies prospectively. An evaluation of a prospective program's impact—rather than the time and money already spent—seems like something that RE researchers would be happy to advocate for.

RE Effects on Policy. It has been said that since RE is "credibly establishing which programs work and which do not, the international agencies can counteract skepticism about the possibility of spending aid effectively and build long-term support for development. Just as randomized trials revolutionized medicine in the twentieth century, [REs] have the possibility to revolutionize social policy during the twenty-first."[42] Moreover, REs are thought to present a simple form of unambiguous evidence that is more likely to influence policy than other evidence connected with empirical development. Here, the great success story is PROGRESA in Mexico, which was scaled up and continued under two different administrations in part because of the positive results of REs.[43]

At the same time, some have doubts about RE's effects on policy, arguing that much of PROGRESA's success in Mexico, for example, was due to political factors, particularly since municipalities that had previously voted for the party in power were more likely to be enrolled in the program, despite attempts to depoliticize it.[44] Even those who dispute that finding have observed that a nondiscretionary PROGRESA/OPORTUNIDADES program paid off at the polls for the incumbent in both the 2000 and 2006 elections, and that President Vicente

40. For example, Ashraf and others (2009) introduce a new savings product and evaluate the impact, and Dupas (2009) introduces a new HIV education program and evaluates its impact.

41. For example, Cohen and Dupas (2009) explore the impact of increasing the prevailing subsidy level for bed nets in Kenya.

42. Duflo and Kremer (2008).

43. Levy (2006).

44. Green (2005).

Fox's decision to expand OPORTUNIDADES from rural areas to the cities made political sense since his party's political base was urban.[45]

In other instances, REs have clearly failed to translate into program adoption, a famous example being the evaluation of private school vouchers in Colombia.[46] Despite the accolades heaped on the program by the RE, it was discontinued and never revived. The RE literature itself may be less interested in influencing policy than RE proponents would like (the cancellation of the Colombian voucher program, for example, receives little mention in this huge literature). Moreover, many results in the literature are based on NGO endeavors, not government projects.

Some would even argue that REs can have a minimal impact on policy at best. In Hausmann's view, set forth in chapter 6, the policy environment is too "complex" to be informed by REs that can inevitably vary only several dimensions at once. In this sense, it is naïve to search for simple policy solutions to development. The issue of policy complexity is particularly problematic when spillover effects are likely to be present. Ravallion (this volume) notes that in most REs the control group cannot really be said to be untreated because limited government/NGO resources and attention are likely to flow into control areas when the treated areas benefit from an intervention. Jessica Cohen (this volume) points out that REs focusing on whether a particular policy or program increases uptake of a public health product are likely to often miss behavioral spillover effects that could ultimately influence the overall morbidity or mortality impact.

Anne Case notes that the political receptivity to outsiders' intervening in social service delivery to the poor is very much an unresolved question. The solution of bypassing the government is also doubtful because of the uneven quality of private service delivery. Case points out that all of this seriously tempers Boone and Johnson's optimism that REs can lead the way to wiping out "pockets of poverty."

Paul Romer also criticizes the naïveté of RE-based policy advice that "fails to take into account what people know about why government policymakers do what they do" and does not recognize that "policymakers are constrained, but rarely ignorant." He predicts that "the experimentalists will overstate their conclusions if they too try to change what practitioners do." He recommends instead that researchers NOT try to be policymakers but just show those who are "how scientific tools like experiments can help them do their jobs."

Using education as an example, Lant Pritchett argues in chapter 5 that government behavior as driven by economists' normative recommendations performs very poorly as a positive model. He also points out that the policy effects of RE cannot themselves be identified using RE: "the randomization agenda as a methodological approach inherits an enormous internal contradiction—that all empirical

45. Diaz-Cayeros and others (2008).
46. Angrist and others (2002).

claims should only be believed when backed by evidence from randomization, excepting, of course, those enormous (and completely unsupported) empirical claims about the impact of randomization on policy."

On the other hand, notes Ben Olken, one cannot ignore the role of evidence in policymaking in general (see his comment on chapter 5). Even though knowledge about "what works" is not guaranteed to change policy, he argues it should more often than not move policymakers in the right direction, in part because they want to get reelected. If politicians can be convinced of the benefits of a particular policy or program for their constituents, they may be willing to adopt it, particularly if the results are presented as rigorous and transparent. Another reason that RE results could be adopted into policy, says Olken, is that they often inform the experts who are called in to consult with governments.

With so much of development policy and programming driven by fads, unproved hypotheses, and anecdotal evidence, a powerful but often overlooked benefit of REs is their ability to highlight not only what works but also what does not. REs can very transparently illustrate that money and effort are flowing in the wrong direction. A 2009 study of bed-net subsidies for pregnant women in Kenya (discussed by Rodrik, this volume) is a good example of an RE that directly influenced policy by illustrating that something of a development "fad" (social marketing) was neither effective nor cost-effective in this case. By illustrating that, contrary to conventional wisdom, usage of public health products given for free need not be lower than those sold for a positive price—and that social marketing funds were being misdirected—the study played an important role in the Kenyan government's decision to make bed nets free for pregnant women.

Even if there is often no direct link between evidence and implementation, it is hard to believe that a robust knowledge of effective development programs is not useful. Knowing what works should bring governments closer to effective programs than chaos and confusion can. This is becoming increasingly apparent as more and more institutions that hold the purse strings of foreign aid are linking funding to evidence. For better or worse, governments tend to have limited power over how foreign aid will be spent in their country. If a donor such as the World Bank or the Gates Foundation decides that it wants to fund a child health program on the basis of evidence generated from the types of REs advocated by Boone and Johnson in chapter 3, for example, recipient country governments are unlikely to refuse. Nancy Birdsall (this volume) discusses an example of how donors might compel governments to adopt programs that acquired evidence suggests are important for development. "Cash on Delivery" (COD) aid can link funds for development directly to the achievement of certain objectives, which could be measures of broad outcomes (such as graduation rates) or of more specific input (for example, student-teacher ratios). Linking aid to some broad objective (and remaining agnostic about how best to accomplish it) is appealing in that it resolves the problem of policy complexity that is raised by Hausmann and others in this volume.[47]

Ethical Concerns. Randomized experiments involve human subjects, which raises tricky ethical issues. Whereas most universities have review boards to address such issues, other bodies that undertake REs may not undergo this kind of scrutiny and thus may be unable to ensure that valuable (possibly life-saving) treatments are not being withheld for scientific research purposes, which is morally objectionable, of course. That such choices are made randomly is of little consolation to those denied the treatment in a community and may even heighten the sense of mystery and distrust surrounding the researchers and what these outsiders are really up to. When the treatment and control groups are in the same or adjacent communities, the RE may generate envy and resentment. RE implementation offers researchers little advice on how to engage local communities to address such issues together.

RE defenders would point out, however, that resources are never unlimited, so it is impossible to treat everybody. Random assignment is a well-accepted device for allocating scarce resources fairly. The RE can be designed in a way to minimize some of the problems just mentioned. Often the treatment is phased in, so the control group in one period becomes the treatment group in the next period. Control and treatment groups can also be spatially separated to minimize envy and resentment. Finally, it is certainly becoming the norm for REs to undergo ethical approval.

Social Engineering and the Lack of a Historical Track Record. What if REs were to succeed in influencing policies on a large scale—would that be a good thing? Needless to say, this question cannot be answered by RE methodology, but more casual empiricism would detect the lack of any obvious examples of countrywide escapes from poverty using policies determined by REs. So the large-scale application of REs to determine policy is an untested bit of social engineering, the outcomes of which would be hard to predict.

As David Weil notes later in this volume, there is insufficient discussion about why outsiders promoting REs to provide social services are doing what they are doing:

> Do economists know something that poor people in developing countries do not know and therefore plan to do something the poor would have done if they were more knowledgeable? Do economists have a different discount rate than they do? Do economists place a different value on the lives of their children or some such thing? Being explicit about why one wants to be in the business of providing social services might help in designing policies that achieve one's goals.

Another real-world consideration noted by Ross Levine is the importance of tacit knowledge for both macro- and micropolicy implementation. One could do

47. Although this raises some new concerns about negative spillover effects on other (nonincentivized) objectives.

rigorous experiments to confirm the physics of hitting a baseball and come up with detailed recommendations based on proven models of physics involving bat speed and location and responding to the speed and trajectory of the baseball. Or one could let a kid practice hitting baseballs for years, which would develop tacit knowledge that cannot be codified into written recommendations. Policy formulation and implementation must surely involve some tacit knowledge that REs cannot supply. It follows also, as Levine notes, that current policies and institutions exist for reasons partly based on such tacit knowledge. Even if they are not optimal, changing policies through RE operating on a presumed blank slate is unlikely to be optimal either.

What do societies do instead of RE? Presumably they rely on some kind of social knowledge and learning to inform their choice of economic institutions and policies. As Banerjee points out in chapter 7, there is little evidence that "growth policy experts" have played any role in making growth happen. Between 1960 and 2008—the period in which economists repeatedly failed to explain growth differences between countries and repeatedly failed in large-scale attempts to raise growth—developing countries grew at the highest rate in human history: 2.7 percent a year per capita, which implies a 3.5 times increase in income. It is also at least anecdotally interesting that East Asia had few academic economists with international reputations during its period of rapid growth, while slow-growing Latin America was awash in them.

Economic arguments still may get some credit. This was also the period in which development policy shifted away from state planning toward more market-friendly approaches, informed in part by big macrofacts, like the failure of planned systems compared with market systems over the long run. It helped that this failure had been predicted by economists like Friedrich Hayek and Milton Friedman even as the planned systems appeared to be doing well. Policy and growth regressions offer no evidence that any part of the growth just cited is due to this shift, but the long-run levels suggest that this shift will pay off sooner or later.

Historically, of course, today's rich countries managed to provide public goods without using REs at all. Alternative mechanisms worked, such as feedback from citizens to politicians through voting, interest group lobbying, decentralization of power to the local level appropriate for most of the public goods, competition between jurisdictions to attract mobile factors, or just "calling your congressman." And, even more obviously, the supply of private goods is not guided by RE testing of "which private goods work," but simply by the choices of consumers in competitive markets. The RE literature could give a little more recognition to these alternative mechanisms for deciding upon public or private goods.

Choice and research evidence need not merely be substitutes, however; they could also be complements. Economic research, both macro and micro, naturally aspires to make more knowledge available to those who make the choices (both leaders and citizens).

Evidence versus Prejudice

REs represent progress in having added to the kit of empirical research a tool that alters the priors of other academics as well as policymakers when there is a strong result (particularly if it helps test a behavioral model). The effect on priors is perhaps the real acid test of what this methodology has to contribute. One commonly cited benefit of REs is that they have raised the bar for what is considered plausible evidence about what works in development. While the highest standard for evidence may not always be an RE, and one may choose to ask questions that cannot be tackled with this methodology, the centrality of REs seems to have made policymakers take issues of endogeneity, selection bias, and sound causal estimation more seriously.

We close with the conciliatory thought that the most relevant divide in development may *not* be between micro and macro or between aggregate data and REs, but rather between those who value objective evidence and those who do not. The development policy discussion has been dominated to an astonishing degree by wishful thinking, baseless assertions, and logical and statistical fallacies. Equally amazing, development efforts keep trying the same thing over and over again, despite a long record of previous failures. By way of example, a computer kiosk program for the poor in India failed to work because of unreliable electricity and Internet connectivity, yet the World Bank's "Empowerment Sourcebook" noted "the success of the initiative." Responding to criticism of the sourcebook, Bank officials stated that the document was merely indicating that the institution intended to help achieve "greater empowerment." One may well ask, as does Banerjee: "Helped to achieve greater empowerment? Through non-working computers?"[48]

Development economists can continue to quarrel about what methodology produces respectable evidence, but they clearly agree on the much larger question of what is *not* respectable evidence, namely, most of what is currently relied on in development policy discussions. What unites us is larger than what divides us.

References

Acemoglu, Daron, Simon Johnson, and James A. Robinson. 2001. "The Colonial Origins of Comparative Development: An Empirical Investigation." *American Economic Review* 91 (December): 1369–1401.

———. 2004. "Institutions as a Fundamental Cause of Long-Run Growth." Working Paper 10481. Cambridge, Mass.: National Bureau of Economic Research.

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

Ashraf, Nava, Dean S. Karlan, and Wesley Yin. 2007. "Female Empowerment: Impact of a Commitment Savings Product in the Philippines." Discussion Paper DP6195. Center for Economic and Policy Research (March) (http:ssrn.com/abstract=1134236).

48. Banerjee (2007, pp. 77, 112).

Baker, Judy L. 2000. *Evaluating the Impact of Development Projects on Poverty: A Handbook for Practitioners.* Washington: LCSPR/PRMPO, World Bank

Banerjee, Abhijit. 2007. *Making Aid Work*. MIT Press.

Burnside, Craig, and David Dollar. 2000. "Aid, Policies, and Growth: Revisiting the Evidence." Policy Discussion Working Paper 3251. Washington: World Bank.

Cohen, Jessica, and Pascaline Dupas. 2009 (forthcoming). "Free Distribution or Cost-Sharing? Evidence from a Randomized Malaria Prevention Experiment." *Quarterly Journal of Economics.*

Collins, Daryl, and others. 2009. *Portfolios of the Poor: How the World's Poor Live on $2 a Day*. Princeton University Press.

Deaton, Angus. 2006. "Global Patterns of Income and Health: Facts, Interpretations, and Policies." Annual Lecture 10. World Institute for Development Economics. Helsinki (September).

———. 2009. "Instruments of Development: Randomization in the Tropics and the Search for the Elusive Keys to Economic Development." Working Paper 14690. Cambridge, Mass.: National Bureau of Economic Research.

Diaz Cayeros, Alberto, Federico Estévez, and Beatriz Magaloni. 2008. "Strategies of Vote Buying: Social Transfers, Democracy and Welfare in Mexico." Stanford Department of Political Science. Draft manuscript.

Duflo, Esther. 2004. "Scaling Up and Evaluation." In *Accelerating Development,* edited by Francois Bourguignon and Boris Pleskovic, pp. 342–67. Oxford University Press.

Duflo, Esther, Michael Kremer, and Jonathan Robinson. 2009. Nudging Farmers to Use Fertilizer: Theory and Experimental Evidence from Kenya, mimeo.

Duflo, Esther, Rachel Glennerster, and Michael Kremer. 2008. "Using Randomization in Development Economics Research: A Toolkit." In *Handbook of Development Economics*, vol. 4, edited by T. Paul Shultz and John Strauss. Amsterdam: North-Holland.

Duflo, Esther, and Michael Kremer. 2008. "Use of Randomization in the Evaluation of Development Effectiveness." In *Reinventing Foreign Aid*, edited by William Easterly, pp. 93–120. MIT Press.

Dupas, Pascaline. 2009. "Do Teenagers Respond to HIV Risk Information? Evidence from a Field Experiment in Kenya." Working Paper 14707. Cambridge, Mass.: National Bureau of Economic Research.

Durlauf, S., P. Johnson, and J. Temple. 2006. "Growth Econometrics," in *Handbook of Economic Growth*, edited by P. Aghion and S. Durlauf. Amsterdam: North-Holland.

Easterly, William. 2001. *The Elusive Quest for Growth: Economists' Adventures and Misadventures in the Tropics.* MIT Press.

Easterly, William, and others. 1993. "Good Policy or Good Luck? Country Growth Performance and Temporary Shocks." *Journal of Monetary Economics* 32, no. 3: 459–83.

Easterly, William, and Ross Levine. 2001. "It's Not Factor Accumulation: Stylized Facts and Growth Models." *World Bank Economic Review* 15, no. 2.

———. 2003. "Tropics, Germs, and Crops: The Role of Endowments in Economic Development." *Journal of Monetary Economics* 50, no. 1 (January): 3–39.

Easterly, William, Ross Levine, and David Roodman. 2004. "New Data, New Doubts: A Comment on Burnside and Dollar's 'Aid, Policies, and Growth' (2000)." *American Economic Review*, 94, no. 3 (June): 774–80.

Fehr, Ernst, and others. 2007. "Fairness and Contract Design." *Econometrica* 75, no. 1: 121–54.

Galor, Oded, and David N. Weil. 2000. "Population, Technology, and Growth: From Malthusian Stagnation to the Demographic Transition and Beyond."*American Economic Review* 90, no. 4: 806–28.

Gilovich, Thomas, Dale Griffin, and Daniel Kahneman. 2002. *Heuristics and Biases: The Psychology of Intuitive Judgment*. Cambridge University Press.

Green, Tina. 2005. "Do Social Transfer Programs Affect Voter Behavior? Evidence from Progresa in Mexico, 1997–2000." University of California, Berkeley. Photocopy.

Harberger, Arnold. 2003. "Sound Policies Can Free Up Natural Forces of Growth." *IMF Survey* 32, no. 13: 213–16.

Imbens, Guido W. 2009. "Better Late than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009)." Working Paper 14896. Cambridge, Mass.: National Bureau of Economic Research.

Imbens, Guido W., and Joshua D. Angrist. 1994. "Identification and Estimation of Local Average Treatment Effects." *Econometrica* 62 (March): 467–75.

Imbens, Guido W., and Jeffrey Woolridge. 2008. "Recent Developments in the Econometrics of Program Evaluation." Working Paper 14251. Cambridge, Mass.: National Bureau of Economic Research.

Jones, Chad. 2005. "Growth and Ideas." In *Handbook of Economic Growth*, edited by Philippe Aghion and Steven Durlauf (www.econ.berkeley.edu/~chad/handbook200.pdf).

Kahneman, Daniel, and Amos Tversky. 2000. *Choices, Values, and Frames*. Cambridge University Press.

Kahneman, Daniel, Paul Slovic, and Amos Tversky, eds. 1982. *Judgment under Uncertainty: Heuristics and Biases.* Cambridge University Press.

Kahneman, Daniel, and Amos Tversky. 1982. "Belief in the Law of Small Numbers." In *Judgment under Uncertainty: Heuristics and Biases*, edited by Daniel Kahneman, Paul Slovic, and Amos Tversky, pp. 23–31. Cambridge University Press.

Klenow, Peter J., and Andrés Rodríguez-Clare. 1997. "Economic Growth: A Review Essay." *Journal of Monetary Economics* 40 (December): pp. 597–617.

Kremer, Michael. 1993. "Population Growth and Technological Change: 1,000,000 B.C. to 1990." *Quarterly Journal of Economics* 108 (August): 681–716.

Kremer, Michael, Jean N. Lee, and Jonathan Robinson. 2008. "The Return to Capital for Small Retailers in Kenya: Evidence from Inventories," unpublished.

Levy, Santiago. 2006. *Progress against Poverty: Sustaining Mexico's Progresa-Oportunidades Program*. Brookings.

Pritchett, Lant. 2008. "It Pays to Be Ignorant: A Simple Political Economy of Rigorous Program Evaluation." In *Reinventing Foreign Aid*, edited by William Easterly, pp. 121–44. MIT Press.

Pritchett, Lant, Deepa Narayan, and Soumya Kapoor. 2009. *Moving Out of Poverty: Success from the Bottom Up*. New York: Palgrave/Macmillan for the World Bank.

Pritchett, Lant, Dani Rodrik, and Ricardo Hausmann. 2005. "Growth Accelerations." *Journal of Economic Growth* 10, no. 4: 303–29.

Rodrik, Dani, A. Subramanian, and F. Trebbi. 2004. "Institutions Rule: The Primacy of Institutions over Geography and Integration in Economic Development." *Journal of Economic Growth* 9, no. 2 (June).

Solow, Robert. 2007. "The Last 50 Years in Growth Theory and the Next 10." *Oxford Review of Economic Policy* 23, no. 1: 3–14.

Werker, Eric D., Faisal Z. Ahmed, and Charles Cohen. 2009. "How Is Foreign Aid Spent? Evidence from a Natural Experiment." *American Economic Journal: Macroeconomics* 1, no. 2 (July): 225–44.

World Bank Commission on Growth and Development. 2008. *The Growth Report: Strategies for Sustained Growth and Inclusive Development.* Washington.