

DOING IMPACT EVALUATION

No.

11

Evaluation in the Practice of Development



THE WORLD BANK

Poverty Reduction and
Economic Management



Thematic Group on Poverty Analysis, Monitoring and Impact Evaluation

Evaluation in the Practice of Development

March 2008

Acknowledgements

This paper was written by Martin Ravallion¹. Helpful comments were provided by Francois Bourguignon, Asli Demirguc-Kunt, Gershon Feder, Jed Friedman, Emanuela Galasso, Markus Goldstein, Bernard Hoekman, Beth King, David McKenzie, Luis Seven, Dominique van de Walle, and Michael Woolcock.

This note was financed by the Trust Fund for Environmentally and Socially Sustainable Development supported by Finland and Norway and by the Bank-Netherlands Partnership Program.

¹ These are the views of the author, and should not be attributed to the World Bank or any affiliated organization.

TABLE OF CONTENTS

INTRODUCTION	1
I. THE UNEASY RELATIONSHIP BETWEEN RESEARCH AND PRACTICE	2
II. WHY MIGHT WE UNDER-INVEST IN EVALUATIVE RESEARCH?	3
III. TOOLS FOR POLICY MAKING AND EVALUATION.....	5
IV. MAKING EVALUATIONS MORE USEFUL FOR PRACTITIONERS.....	7
V. UNDERSTANDING IMPACT.....	14
VI. PUBLICATION BIASES.....	20
CONCLUSIONS.....	22
REFERENCES	23

Introduction

Anyone who doubts the potential benefits to development practitioners from evaluation should study China's experience at economic reform. 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" (Du Runsheng, 2006, p.2). 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 (Xiaopeng Luo, 2007). 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.

Unfortunately we still have a long way to go before we will be able to say that this story from China is typical of development policy making elsewhere. Practitioners—working in developing countries, donor countries and international agencies—search continually for operational solutions to pressing development problems. Researchers have the training and skills needed to provide the conceptual and technical tools to inform that search, and help learn from our success and failures along the way. However, practitioners rarely appreciate the full benefits from research. And, as a rule, it is the development practitioners who hold the purse strings, or have greater influence on, public spending on evaluation. The result is that too little rigorous research of relevance to development gets done. We know too little about the efficacy of development efforts and the learning process becomes too weak to reliably guide practice. The outcome is almost certainly less overall impact on poverty.

How can we assure that the potential gains from using research in development practice are realized? This requires a type of research that I will call "evaluative research," which tries to probe deeply into the lessons from past policies, and so contribute to debates about future policies. This is research that rigorously assesses the benefits and costs of development policies and under what circumstances the benefits might be higher, or the costs lower. It is done both *ex ante* (before the intervention) and *ex post* (after it). Evaluative research goes beyond telling us whether a specific policy or program delivered its intended outcomes *ex post*, but aims to critically expose the conceptual and empirical foundations for policy making *ex ante*, as well as learning why some efforts succeeded and some failed.

I will argue that too little evaluative research on development effectiveness gets done and that the evaluations that are done currently are not as useful as they could be for

learning about development effectiveness. The production of evaluative research is riddled with problems of externalities, selection-biases in funding, myopia, publication biases and weak incentives to innovate in developing and adapting the tools needed. I shall point to a number of things that need to change if the potential for evaluative research is to be fulfilled. These include addressing the problems that lead to suboptimal investment in research, as well as things that researchers need to do to make their work more relevant to the needs of practitioners.

I. The Uneasy Relationship between Research and Practice

Academic researchers draw the bulk of their ideas from the work of other academic researchers. To be useful, evaluative research needs to draw more heavily on inputs from non-researchers actively involved in making and thinking about policy in developing countries. However, while the relationship between research and practice is important for development effectiveness, it is not a particularly easy relationship for either side. Three generic sources of tension in the relationship can be identified.

The first source of tension in the relationship concerns the role of advocacy. Practitioners often take on an advocacy role, as they strive to improve peoples' lives. This can be in tension with the perspective of researchers who are more inclined to critically question the foundations of any policy position—to see if it really does improve peoples' lives. Balancing advocacy and rigor can be difficult, although it is usually clear when one has gone too far in one direction. Even when research findings are properly presented (with all the caveats acknowledged), practitioners keen to demonstrate impact are sometimes tempted to highlight research findings selectively, or to engage with researchers selectively, such as when practitioners only put forward their best projects for serious evaluation. Sometimes researchers have to point out such biases, and doing so does not always make them popular with practitioners.

It must also be acknowledged that there is a danger that researchers, eager for impact, overreach in their efforts to have impact on public debates. This is not helped by the tendency of the research community to (on the one hand) encourage serious effort to assure the internal validity of a piece of research within its predetermined scope, but (on the other hand) to be quite relaxed about the claims made about implications for policy or relevance to other settings. A few researchers have also been swayed by the publicity, and even fame, that can come when high-profile advocacy is disguised as research. This can undermine the credibility of the researcher, and can also spillover to jeopardize the credibility of other researchers. The long-term interests of development are not well served by researchers turning into advocates.

A second source of tension concerns the role played by "policy rules." It is not always practical to adapt development policies to each context. Understandably, practitioners look for "best-practice" generalizations to guide action in diverse settings with limited information. This often runs into conflict with a researcher's perspective, which emphasizes the uncertainties and contingent factors in policy choices and points to

the importance of context to outcomes and the need for better data. (Though, as I have noted, researchers are sometimes too hasty in drawing lessons beyond the scope of their study.) In principle one might develop a complete mapping from every possible circumstance to every policy, but that is rarely feasible in reality. So the tension is inevitable between practitioners looking (often in vain) for robust rules to guide action and researchers pointing to the caveats and unknowns.

The role of advocacy and the practical need for policy rules entail that, at any one time, there is a received wisdom about development policy. Evaluative research is the main way in which such received wisdoms come to be challenged and, when needed, revised. For example, advocates of free trade have asserted that it would dramatically reduce absolute poverty in the world; research has provided a more qualified view based on evaluations of specific trade reforms (Hertel and Winters, 2006; Ravallion, 2006; Hoekman and Olarreaga, 2007).

The third reason why researchers and practitioners do not always see eye-to-eye is that they often have different time frames. Practitioners working on specific projects or policy reforms need rapid feedback. Researchers need time, both to collect the required data and analyze it. Impatient practitioners are swayed by various short-cut methods of “impact assessment” that offer quick results of doubtful or unknown veracity.

Related to this difference, researchers are sometimes drawn to “blue-sky” topics. Not every relevant question that should be addressed by researchers is currently being asked by practitioners. There are examples in which researchers have explored issues that were not being asked by practitioners at the time, but became relevant later. For example, the current recognition of the potential for land titling or “regularization” as a viable development intervention (and the basis of numerous lending operations by the World Bank since the 1990s) emerged out of research on the economic benefits of greater security of landholdings; an influential research study was Feder et al. (1988). To give another example, researchers at the World Bank had pointed to concerns about whether the deposit insurance arrangements being advocated by practitioners were appropriate in countries with poorly developed financial institutions and weak regulatory environments (Demirguc-Kunt and Detragiache, 2002). The research findings eventually influenced Bank operations. Carefully-chosen blue-sky topics, including methodological research (that I return to below), can have large pay offs; foresight is often needed to make development research relevant.

II. Why Might We Under-invest in Evaluative Research?

Rigorous evaluations tied directly to development practice are rarely easy. Practical and logistical difficulties abound in designing and implementing sound evaluations. Special-purpose data collection and close supervision are typically required. The analytic and computational demands for valid inferences can also be daunting and require specialized skills. Practitioners cannot easily assess the expected benefits. Short-cut methods promise quick results at low cost, though rarely are users well informed of

the inferential dangers.² Information asymmetries entail that rigorous evaluations are driven out by non-rigorous ones.

Evaluative research also has some of the properties of a public good, in that the benefits spillover to other projects. Development is a learning process, in which future practitioners benefit from current research. The individual project manager will typically not take account of these external benefits when deciding how much to spend on research. The decision about whether resources should be invested in data and research on a specific project or policy is often made by (or heavily influenced by) the individual practitioners involved. Yet a significant share of the benefit accrues to others, including future practitioners. This is what economists call an “externality.” It is an important reason why research needs to be supported institutionally—supplementing the funding that is provided by the specific interventions under study.

Certain types of evaluative research are likely to be more prone to this problem. It is typically far easier to evaluate an intervention that yields all its likely impact within one year than an intervention that takes many years. It can be no surprise that credible evaluations of the longer-term impacts of (for example) infrastructure projects are rare. Similarly, we know very little about the long-term impacts of development projects that do deliver short-term gains; for example, we know much more about the short-term impacts of transfers on the current nutritional status of children in recipient families than about the possible gains in their longer-term productivity from better nutrition in childhood. So future practitioners are poorly informed about what works and what does not. There is a “myopia bias” in our knowledge, favoring development projects that yield quick results.

We probably also under-invest in evaluative research on types of interventions that tend to have diffused, wide-spread, benefits. Impacts for such interventions are often harder to identify than for cleanly assigned programs with well-defined beneficiaries, since one typically does not have the informational advantage of being able to observe non-participants (as the basis for inferring the counterfactual). It may also be harder to fund evaluations for such interventions, since they often lack a well-defined constituency of political support.

The implication is that, without strong institutional support and encouragement, there will probably be too little evaluative research and, in particular, too little research on long-term impacts of development interventions and of broader sectoral or economy-wide reforms. The fact that long-term evaluations are so rare (yet it is widely agreed that development does not happen rapidly) suggests that the available support is insufficient.

² For example, OECD (2007) outlines an approach to “*ex ante* poverty impact assessment” that claims to assess the “poverty outcomes and impacts” of an aid project in just 2-3 weeks at a cost of \$10,000-\$40,000 (even less if done as “an integral part of the appraisal process”). Essentially a consultant fills in a series of tables giving the project’s “short-term and long-term outcomes” across a range of (economic and non-economic) dimensions for each of various groups of identified “stakeholders,” as well as the project’s “transmission channels” through induced changes in prices, employment, transfers and so on. Virtually no indication is given to readers of just how difficult it is to make such assessments in a credible way, and what degree of confidence one can have in the results.

Increasingly evaluation does receive support beyond what is demanded by the immediate practitioners. There has been substantial growth in donor support for impact evaluations in recent years. Donor governments are being increasingly pressed by their citizens to show the impact of development aid, which has meant extra resources for financing impact evaluations. While the resources available do not always go to rigorous evaluative research, researchers have helped stimulate broader awareness of the problems faced when trying to do evaluations, including the age-old problem of identifying causal impacts. This has helped make donors less willing to fund weak proposals for evaluations that are unlikely to yield reliable knowledge about development effectiveness.

Nonetheless, what gets evaluated is still a modest fraction of what gets done on the ground in the name of development. More worrying though is that it is a decidedly non-random fraction. On top of the biases due to asymmetric information and myopia, external funding is almost entirely demand driven, and no less so now, even with the higher levels of funding. A self-selected sample of practitioners approaches the funding sources, often with researchers already in tow. This process is likely to favor projects and policies that are expected to have benefits by their advocates.

III. Tools for Policy Making and Evaluation

Practitioners carry a set of tools—data, measures and models—to their project and policy dialogues with governments. Researchers can usefully assess those tools and help improve them. Are the data sound? Are the measures the most relevant ones to the specific policy problem? Are the assumptions both realistic and internally consistent?

The “policy rules” discussed above are examples of the tools that practitioners bring to policy dialogues, and researchers play an important role in continually questioning those rules. To give one example, policy-oriented discussions often assume that “better targeting” of direct interventions against poverty (such as cash transfer schemes) implies larger impacts on poverty or more cost-effective interventions for fighting poverty. The literature on the economics of targeting has warned against that assumption, but evidence has been scarce, and the lessons from the literature have often been ignored by practitioners.³ There appears to be much scope for critically assessing the measures and rules used by development practitioners.

Data are also undeniably important tools for practitioners. The bulk of the data that practitioners use is pre-existing; practitioners are typically drawing on public goods generated by prior data-collection efforts. So issues about incentives and financing are crucial to the production of policy-relevant data. Here again, the public good nature of data is an issue. Public access data will almost certainly be undersupplied by individuals working on their own.

³ In one case study, it was demonstrated that the standard measures of targeting performance used by practitioners are quite uninformative, or even deceptive, about the impacts on poverty, and cost-effectiveness in reducing poverty, of a large cash transfer program in China (Ravallion, 2007).

There are other reasons why we under-invest in the tools of evaluative research. Tool development often entails bridging gaps between theory and practice, including developing tools appropriate to practice. Such work is not typically valued highly in academia, where greater emphasis is given to theoretical and deeper methodological developments. At the same time, while tool development has received support from development aid agencies, higher priority tends to be given to efforts with more immediate bearing on specific policy issues. Efforts focused directly on tool development for practitioners are not academic enough by some perceptions, but too academic in the eyes of others, including those who provide funding.

While development data still remain weak, we can point to progress in some areas. Much of the data now used routinely are the products of past research projects, and it would seem very unlikely that any of these data would exist without strong initial public support, given the extent of the externalities problem. A number of the World Bank's most successful data initiatives started as research projects.⁴ The *Living Standards Measurement Study* (LSMS) is one example (see <http://www.worldbank.org/lsms/>). This started in the Bank as a research project around 1980, with the aim of greatly improving household survey data in developing countries; the LSMS has now done over 80 large multi-purpose household surveys in 50 countries and has set the standard for survey data collection in developing countries. At a more macro level, international trade data and methods provide an important class of examples of global public goods in which tool development would be unlikely to happen if one relied solely on development practitioners at country level (both individual governments and the "country teams" of international agencies). Other examples include the various cross-country data bases that have been developed out of research projects aiming to assess country performance in both economic, political and social domains. The Bank's global poverty monitoring effort that started as a small research project for a background paper to the 1990 *World Development Report*.⁵ The Bank's *Enterprise Surveys* (formally known as *Investment Climate Surveys*) started in its research department, but were eventually mainstreamed in the relevant sectoral units. These are global public goods that would never have happened without initiatives by researchers and strong institutional support. Indeed, I doubt if there would have been any of this progress without strong institutional support—going well beyond the immediate needs of individual practitioners.

However, research could do more to help expand the tool kit routinely employed by policy makers and analysts in deciding what policies and programs to implement and how to assess their performance. Theoretical work often points to new tools, some more useful than others. Choosing wisely is important. Researchers can have great value to practitioners in demonstrating the usefulness of new analytic tools in real applications. Over the years, the Bank's researchers have developed numerous software programs that have facilitated data analysis throughout the Bank and in client countries.⁶

⁴ I focus here on World Bank examples; one could also mention other important example such as USAID's *Demographic and Health Surveys* (<http://www.measuredhs.com/>) and RAND's *Family Life Surveys*.

⁵ See World Bank (1990); the background paper was Ravallion et al. (1991).

⁶ See, for example, the "Poverty Analysis Toolkit" produced by the Bank's research department (<http://econ.worldbank.org/programs/poverty>).

There is also much to be done in making data-collection efforts (including administrative data) more relevant as evaluative tools. There has been a longstanding need to develop better survey modules on program participation and access to public facilities. The design of routine surveys could also be more useful for evaluative research. For example, by allowing only partial rotation of samples at each survey round one can create longitudinal observations that can facilitate the construction of baseline observations for impact evaluations.⁷

IV. Making Evaluations More Useful for Practitioners

The archetypal formulation of the evaluation problem aims to estimate the average impact on those to which a specific program is assigned (the participants) by attempting to infer the counterfactual from those to which it is not assigned (non-participants). While this is an undeniably important and challenging problem, solving it is not sufficient for assuring that evaluation is relevant to development practice. The box outlines 10 steps toward making evaluation more relevant. It is acknowledged that there are some difficult problems ahead in following these steps, and that more research is needed; the main aim here is to point in useful directions for future work.

Box 1 : Ten Steps to Making Impact Evaluations More Relevant

1. *Start with policy-relevant questions.* The question typically asked is: did the intervention have an impact? It is often important to also ask: why is the intervention needed and why should one expect it to have an impact? The most important questions may not even involve a specific intervention, but rather an issue relevant to the case for intervention; for example, how well do markets work in the absence of intervention? Be wary of constraining evaluative research to situations in which some researcher's favorite method is feasible; this may well exclude important and pressing development questions.
2. *Take the ethical objections and political sensitivities seriously; policy makers do!* There can be ethical concerns with deliberately denying a program to those who need it and providing the program to some who do not; this applies to both experimental and non-experimental methods. For example, with too few resources to go around, randomization may be seen as a fair solution, possibly after conditioning on observables. However, the information available to the evaluator (for conditioning) is typically a partial subset of the information available "on the ground" (including to voters). The idea of "intention-to-treat" helps alleviate these concerns; one has a randomize assignment, but anyone is free to not participate. But even then, the "randomized out" group may include people in great need. All these issues must be discussed openly, and weighed against the (potentially large) longer-term welfare gains from better information for public decision making.
3. *Taking a comprehensive approach to the sources of bias.* Selection bias can stem from both observables and unobservables (to the evaluator), i.e., participants have latent attributes that yield higher/lower outcomes. Some economists have focused on the latter bias, ignoring enumerable other biases/problems in impact evaluation. Less than ideal methods have been used to control for observable heterogeneity including *ad hoc* models of outcomes. There is

⁷ For example, the evaluation by Galasso and Ravallion (2004) of the government's safety net response to the macroeconomic crisis in Argentina exploited this property of the labor force survey; otherwise, no baseline data would have been possible given that one could not (of course) delay the intervention to collect baseline data.

some evidence that we have given too little attention to the problem of selection bias based on observables. In practice, we often find arbitrary preferences for one conditional independence assumption (exclusion restrictions) over another (conditional exogeneity). One cannot scientifically judge the appropriate assumptions/methods independently of the program, setting and data.

4. *Do a better job on spillover effects.* Classic impact evaluations assume that there are no impacts for non-participants, but this is unlikely to hold for many development projects. Spillover effects can stem from market responses (given that participants and non-participants trade in the same markets), the (non-market) behavior of participants/non-participants or the behavior of intervening agents (governmental/NGO). For example, aid projects often target local areas, assuming that the local government will not respond; yet if one village gets the project, the local government may well cut its spending on that village, and move to the control village. We will probably underestimate the impact of the project if we ignore such spillover effects.
5. *Take a sectoral approach.* Given that all development aid is to some extent fungible, you are not always evaluating what the extra public resources actually financed, but rather what the aid recipient chose to put forward. So the evaluation may be deceptive about the true impact of the external aid. A broader approach to evaluation is needed—looking at other activities in the sector and possibly beyond.
6. *Fully explore impact heterogeneity.* Impacts will vary with participant characteristics (including those not observed by the evaluator) and context. Participant heterogeneity implies the need for interaction effects in modeling impacts. There is also heterogeneity in unobserved determinants of impacts plus participant responses to these latent impact factors. This has implications for evaluation methods, project design and external validity of the results. Contextual heterogeneity is plausible in development settings. Local institutional factors can be important to development impact.
7. *Take “scaling up” seriously.* With scaling up one may find that the inputs change (entry effects: nature and composition of those who “sign up” changes with scale; migration responses), the intervention may also change (resource effects) and outcomes may change, such as due to lags in outcome responses, market responses (partial equilibrium assumptions are fine for a pilot but not when scaled up), social effects/political economy effects; early vs. late capture. The “scaled-up” program may be fundamentally different to the pilot. There has been too little work on external validity and scaling up.
8. *Understand what determines impact.* Replication across differing contexts can help, though the feasibility of spanning the relevant range of contexts is unclear. Intermediate indicators can help understand impacts as can qualitative research/mixed methods. It can be used to test the assumptions made in rationalizing an intervention. In understanding impact Step 9 is key:
9. *Don’t reject theory and structural modeling.* Standard evaluations are “black boxes”: they give policy effects in specific settings but not structural parameters (as relevant to other settings). Structural methods allow us to simulate changes in program design or setting. However, assumptions are needed; that is the role of theory.
10. *Develop capabilities for evaluation within developing countries.* Strive for a culture of evidence-based evaluation practice. Evaluation is often a natural addition to the roles of the government’s sample survey unit. Independence should already be in place; connectivity to other public agencies may be a bigger problem. Sometimes a private-sector evaluation capability will be required.

Questions for evaluative research. Evaluative research should not take the intervention as pre-determined, but must begin by probing the problem that a policy or project is addressing. Why is the intervention needed? How does it relate to overall

development goals, such as poverty reduction? What are the market, or governmental, failures it addresses? What are its distributional goals? What are the trade-offs with alternative (including existing) policies or programs? Researchers can often play an important role in addressing these questions. This involves more precise identification of the policy objectives (properly weighing gains across different sub-groups of a population, and different generations), the relevant constraints, which include resources, information, incentives and political economy constraints, and the causal links through which the specific intervention yields its expected outcomes.

This role for evaluative research in conceptualizing the case for intervention can be especially important when the capacity for development policy making is weak, or when it is captured by lobby groups, advocating narrow sectoral interests. The *ex ante* evaluative role for research can also be crucial when practitioners have overly strong priors about what needs to be done. Over time, some practitioners become experts at specific types of interventions, and some may even lobby for those interventions. The key questions about whether the intervention is appropriate in the specific setting may not even get asked.

Evaluators themselves can also become lobbyists for their favorite methods. While it may seem trivially true that one should start all evaluative research with interesting and important questions, far too often it is not the question that is driving the research agenda but a preference for certain types of data or certain methods; the question is then found that fits the methodology, not the other way round. Consider, for example, randomized experiments. These can be valuable in helping to identify the causal impacts of interventions; a purely random assignment of who participates (though rare in practice) eliminates the thorny problem of “selection bias” whereby those who chose to participate (or are chosen by the program) have different values of the relevant counterfactual outcomes to those who do not. However, it is often hard to convince a government to randomize a development program, which can be antithetical to its aims; for example, a program that aims to reduce poverty will (hopefully) focus on poor people, not some random sub-sample of people, some of whom are poor and some not. It can be hard to justify to the public why some people in obvious need are deliberately denied access to the program (to form the control group) in favor of some who don’t need it. Even when the randomization is done after allowing for purposive selection based on observables (“conditional randomization”), what is observable to the evaluator or government is typically only a sub-set of what is observable on the ground. Key stakeholders may see quite plainly that the program is being withheld from some people who need it while it is being given to some who do not.

Starting with the question, not the method, often points the researcher toward types of data and methods outside the domain traditionally favored by the researcher’s own disciplinary background. For example, some of the Bank’s research economists trying to understand persistent poverty and the impacts of antipoverty programs have been drawn into the theories and methods favored in other social sciences such as anthropology, sociology and social psychology; see, for example, the collection of papers in Rao and Walton (2004). Good researchers, like good detectives, assemble and interpret diverse forms of evidence in testing empirical claims.

As is now widely appreciated, evaluative research should assess impacts against explicit and relevant counterfactuals, such as the absence of the policy in question or some policy option. This requires sound evaluation designs, using good data and credible strategies for identifying causal impacts from those data—taking proper account of the likely sources of bias, such as when outcomes are only compared over time for program participants, or when participants and non-participants are compared at only one date. This is all about “internal validity,” which has been the main focus of researchers working on evaluations. This discussion will flag some issues that have received less attention and matter greatly to the impact of evaluative research.

The choice of counterfactual is one such issue. The classic evaluation focuses on counterfactual outcomes in the absence of the program. This counterfactual may fall well short of addressing the concerns of policy makers. The alternative of interest to policy makers is rarely to do nothing, but rather to spend the same resources on some other program (possibly a different version of the same program). A specific program may appear to perform well against the option of doing nothing, but poorly against some feasible alternative. For example, in an impact evaluation of a workfare program in India, Ravallion and Datt (1995) showed that the program substantially reduced poverty among the participants relative to the counterfactual of “no program,” but that once the costs of the program were factored in (including the foregone income of workfare participants), the alternative counterfactual of a uniform (un-targeted) allocation of the same budget outlay would have had more impact on poverty. Formally, the evaluation problem is essentially no different if some alternative program is the counterfactual; in principle we can repeat the analysis relative to the “do nothing counterfactual” for each possible alternative and compare them. But this is rare in practice.

Nor is it evident that the classic formulation of the impact evaluation problem yields the most relevant impact parameters. For example, there is often interest in better understanding the horizontal impacts of a program: the differences in impacts at a given level of counterfactual outcomes, as revealed by the joint distribution of outcomes under treatment and outcomes under the counterfactual. We cannot know this from a randomized evaluation, which only reveals net counterfactual mean outcomes for those treated. Instead of focusing solely on the net gains to the poor (say) we may ask how many losers there are among the poor, and how many gainers.

Counterfactual analysis of the joint distribution of outcomes over time is useful for understanding impacts on poverty dynamics. This approach is developed in Ravallion et al. (1995) for the purpose of measuring the impacts of changes in social spending on the inter-temporal joint distribution of income. Instead of only measuring the impact on poverty (the marginal distribution of income) the authors exploit panel data to distinguish impacts on the number of people who escape poverty over time (the “promotion” role of a safety net) from impacts on the number who fall into poverty (the “protection” role). Ravallion et al. apply this approach to an assessment of the impact on poverty transitions of reforms in Hungary’s social safety net.

Spillover effects. A further way in which the classic impact evaluation problem often needs to be adapted to the needs of practitioners concerns its assumption that

impacts for direct participants do not spillover to non-participants; only under this assumption can infer the counterfactual from an appropriate sample of the non-participants. Spillover effects are recognized as a concern in evaluating large public programs, for which contamination of the control group can be hard to avoid due to the responses of markets and governments, and in drawing lessons for scaling up based on randomized trials (Mofitt, 2003).

An example of spillover effects can be found in the Miguel and Kremer (2004) study of treatments for intestinal worms in children. The authors argue that a randomized design, in which some children are treated and some are retained as controls, would seriously underestimate the gains from treatment by ignoring the externalities between treated and “control” children. The design for the authors’ own evaluation avoided this problem by using mass treatment at the school level instead of individual treatment (using control schools at sufficient distance from treatment schools).

Spillover effects can also arise from the way markets respond to an intervention. Ravallion (2008) discusses the example of an *Employment Guarantee Scheme* (EGS) in which the government commits to give work to anyone who wants it at a stipulated wage rate; this was the aim of the famous EGS in the Indian state of Maharashtra and in 2006 the Government of India implemented a national version of this scheme. The attractions of an EGS as a safety net stem from the fact that access to the program is universal (anyone who wants help can get it) but that all participants must work to obtain benefits and at a wage rate that is considered low in the specific context. The universality of access means that the scheme can provide effective insurance against risk. The work requirement at a low wage rate is taken by proponents to imply that the scheme will be self-targeted to the income poor.

The EGS is an assigned program, in that there are well-defined “participants” and “non-participants.” And at first glance it might seem appropriate to collect data on both groups and compare their outcomes either by random assignment or after cleaning out observable heterogeneity. However, this classic evaluation design could give a severely biased result. The gains from such a program are very likely to spillover into the private labor market. If the employment guarantee is effective then the scheme will establish a firm lower bound to the entire wage distribution—assuming that no able-bodied worker would accept non-EGS work at any wage rate below the EGS wage. So even if one picks a perfect comparison group, one will conclude that the scheme has no impact, since wages will be the same for participants and non-participants. But that would entirely miss the impact, which could be large for both groups.

Spillover effects can also arise from the behavior of governments. Chen et al. (2007) find evidence of such spillover effects in their evaluation of a World Bank supported poor-area development program in rural China. When the program selected certain villages to participate, the local government withdrew some of its own spending on development projects in those villages, in favor of non-program villages—the same set of villages from which the comparison group was drawn. Ignoring these spillover effects generated a non-negligible underestimation of the impact of the program. Chen et

al. show how one can estimate the maximum bias due to the specific type of spillover effects that arises from local government spending responses to external development aid.

Heterogeneity. Practitioners should never be happy with an evaluation that assumes common (homogeneous) impact. The impact of an assigned intervention can vary across those receiving it. Even with a constant benefit level, eligibility criteria entail differential costs on participants. For example, the foregone labor earnings incurred by participants in workfare or conditional cash transfer schemes (via the loss of earnings from child labor) will vary according to skills and local labor-market conditions.

Recognizing the scope for heterogeneity in impacts and the role of contextual factors makes evaluative research more relevant to good policy making. For example, in their evaluation of a poor-area development program in rural China, Chen et al. (2007) found low overall impact, but considerable heterogeneity, in that different types of households benefited more than others. The policy implication is that choosing different beneficiaries would have greatly increased the project's overall impact. By developing a deeper understanding of the heterogeneity of impacts researchers will be in a better position to assess the external validity of evaluation findings to other settings.

Heterogeneity of impacts in terms of observables is readily allowed for by adding interaction effects between the intervention and observables to one's model of outcomes. With some extra effort, one can also allow for latent heterogeneity in the impacts of an intervention, using a random coefficients estimator in which the impact estimate contains a stochastic component. Applying this type of estimator to the evaluation data for *PROGRESA* (a conditional cash transfer program in Mexico), Djebbari and Smith (2005) found that they can convincingly reject the assumption of common (homogeneous) effects made by past evaluations of that program.

When there is such heterogeneity, one will often want to distinguish marginal impacts from average impacts. Following Björklund and Moffitt (1987), the marginal treatment effect can be defined as the mean gain to units that are indifferent between participating or not. This requires that we model explicitly the choice problem facing participants (Björklund and Moffitt, 1987; Heckman and Navarro-Lozano, 2004). We may also want to estimate the joint distribution of outcomes under treatment and outcomes under the counterfactual, and a method for doing so is outlined in Heckman et al. (1997).

External validity. Arguably the most important things to learn from any evaluation relate to its lessons for future policies (including possible reforms to actual interventions being evaluated). Although external validity is highly desirable for evaluative research, it can be hard to achieve. We naturally want research findings to have a degree of generalizability, so they can provide useful knowledge to guide practice in other settings. Thus empirical researchers need to focus on questions that tell us about why a policy or program has impact; I return to this question below. However, too often impact evaluations are a "black box"; under certain assumptions, they reveal average impacts among those who receive a program, but say little or nothing about the economic and social processes leading to that impact. And only by understanding those processes

can we draw valid lessons for scaling up, including expanding to other settings. Research that tests the theories that underlie the rationales for policy or program intervention can thus be useful in practice.

When the policy issue is whether to expand a given program at the margin, the classic estimator of mean-impact on the treated is actually of rather limited interest. For example, we may want to know the marginal impact of a greater duration of exposure to the program. An example can be found in the study by Ravallion et al. (2005) of the impacts on workfare participants of leaving the program, relative to staying (recognizing that this entails a non-random selection process). Another example can be found in the study by Behrman et al. (2004) of the impacts on children's cognitive skills and health status of longer exposure to a preschool program in Bolivia. The authors provide an estimate of the marginal impact of higher program duration by comparing the cumulative effects of different durations using a matching estimator. In such cases, selection into the program is not an issue, and we do not even need data on units who never participated.

Relatedly, sound evaluative research must recognize the importance of context since this can be key to drawing valid lessons for other settings. Unless we understand the role of context, the research will have weak external validity. Contextual factors include the circumstances of participants, the cultural and political environment, and the administrative context.

Given that we can expect in general that any intervention will have heterogeneous impacts—some participants gain more than others—there are some serious concerns about the external validity (generalizability) of randomized trials. The people who are normally attracted to a program, taking account of the expected benefits and costs to them personally, may differ systematically from the random sample of people who got the trial.⁸ As already noted, not all sources of heterogeneity are observable, and participants and stakeholders often react to factors unobserved by the researcher, confounding efforts to identify true impacts using standard methods, including randomized experiments; for further discussion see Heckman et al. (2006).

External validity concerns about impact evaluations can also arise when certain institutions need to be present to even facilitate the evaluations. For example, when randomized trials are tied to the activities of specific Non-Governmental Organizations (NGOs) as the facilitators, there is a concern that the same intervention at national scale may have a very different impact in places without the NGO. Making sure that the control group areas also have the NGO can help, but even then we cannot rule out interaction effects between the NGO's activities and the intervention. In other words, the effect of the NGO may not be “additive” but “multiplicative,” such that the difference between measured outcomes for the treatment and control groups does not reveal the impact in the absence of the NGO. Furthermore, the very nature of the intervention may change when it is implemented by a government rather than an NGO. This may happen because of unavoidable differences in (*inter alia*) the quality of supervision, the incentives facing service providers, and administrative capacity.

⁸ This is sometimes called “randomization bias”; see Heckman and Smith (1995). Also see Moffitt (2004).

A further external validity concern is that, while partial equilibrium assumptions may be fine for a pilot, general equilibrium effects (sometimes called “feedback” or “macro” effects) can be important when the pilot is scaled up nationally. For example, an estimate of the impact on schooling of a tuition subsidy based on a randomized trial (such as for Mexico’s *PROGRESA* program) may be deceptive when scaled up, given that the structure of returns to schooling will alter. Heckman et al., (1998) demonstrate that partial equilibrium analysis can greatly overestimate the impact of a tuition subsidy once relative wages adjust, although Lee (2005) finds a much smaller difference between the general and partial equilibrium effects of a tuition subsidy in a slightly different model.

A special case of the general problem of external validity is scaling up. There are many things that can change when a pilot program is scaled up: the inputs to the intervention can change, the outcomes can change, and the intervention can change; Moffitt (2006) gives examples in the context of education programs. The realized impacts on scaling up can differ from the trial results (whether randomized or not) because the socio-economic composition of program participation varies with scale. Ravallion (2004) discusses how this can happen in theory, and presents results from a series of country case studies, all of which suggest that the incidence of program benefits becomes more pro-poor with scaling up. Trial results could over or under estimate impacts on scaling up. Larger projects may be more susceptible to rent seeking or corruption (as Deaton, 2006, suggests); alternatively, the political economy may entail that the initial benefits tend to be captured more by the non-poor (Lanjouw and Ravallion, 1999).

Evaluative research should regularly test the assumptions made in operational work. Even field-hardened practitioners do what they do on the basis of some implicit model of how the world works, which rationalizes what they do, and how their development project is expected to have impact. Existing methods of rapid *ex ante* impact assessment evidently also rely heavily on the models held by practitioners. Researchers can perform a valuable role in helping to make those models explicit and (where possible) helping to assess their veracity.

A case in point is the assumption routinely made by both project staff and evaluators that the donor’s money is actually financing what recipients claim it is financing. Research has pointed to a degree of fungibility in development aid. Donors probably do not fund exactly what aid recipients claim they have funded, although there is some evidence that external aid sticks to its sector; see van de Walle and Mu (2007). The existence of fungibility and flypaper effects points to the need for a broader sectoral approach in efforts to evaluate the impacts of development aid.

V. Understanding Impact

The above discussion points to the need to supplement standard evaluation tools by other sources of information that can throw light on the factors that influence the measured outcomes. That can be crucial for drawing useful policy lessons from the evaluation, including lessons for redesigning a program and scaling it up (including implementation in different settings). The relevant factors relate to both the participants

(such as understanding program take-up decisions and how the outcomes are influenced by participants' characteristics) and program context (such as understanding how the quantity/quality of service provision affects outcomes and the role of local institutions in influencing outcomes).

An obvious approach to understanding what factors influence a program's performance is to repeat it across different types of participants and in different contexts. Duflo and Kremer (2005) and Banerjee (2007) have argued that repeated randomized trials across varying contexts and scales should be used to decide what works and what does not in development aid; Banerjee goes so far as to argue that projects should only be funded when they have had positive results in experimental trials in the relevant settings. Even putting aside the aforementioned problems of social experiments, 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.

Even if one cannot go as far as Banerjee would like, evaluation designs should plan for contextual variation. Important clues to understanding impacts can often be found in the geographic differences in impacts. These can stem from geographic differences in relevant population characteristics or from deeper location effects, such as agro-climatic differences and differences in local institutions (such as local “social capital” or the effectiveness of local public agencies). An example can be found in the study by Galasso and Ravallion in which the targeting performance of Bangladesh's *Food-for-Education* program was assessed across each of 100 villages in Bangladesh and the results were correlated with characteristics of those villages. The authors found that the revealed differences in performance were partly explicable in terms of observable village characteristics, such as the extent of intra-village inequality (with more unequal villages being less effective in reaching their poor through the program). Failure to allow for such location differences has been identified as a serious weakness in past evaluations; see for example the comments by Moffitt (2003) on trials of welfare reforms in the US.

The literature suggests that location is a key dimension of context. An implication is that it is less problematic to scale-up from a pilot within the same geographic setting (with a given set of relevant institutions) than to extrapolate the trial to a different setting. In one of the few attempts to test how well evaluation results from one location can be extrapolated to another location, Attanasio et al. (2003) divided the seven states of Mexico in which the *PROGRESA* evaluation was done into two groups. They found that results from one group had poor predictive power for assessing likely impacts in the other group. Again, the ability to understand location effects on program performance is key to the implications for scaling up.

Useful clues for understanding impacts can sometimes be found by studying what can be called “intermediate” outcome measures. The typical evaluation design identifies a small number of “final outcome” indicators, and aims to assess the program's impact on those indicators. Instead of using only final outcome indicators, one may choose to also

study impacts on certain intermediate indicators of behavior. For example, the inter-temporal behavioral responses of participants in anti-poverty programs are of obvious relevance to understanding their impacts. An impact evaluation of a program of compensatory cash transfers to Mexican farmers found that the transfers were partly invested, with second-round effects on future incomes (Sadoulet et al., 2001). Similarly, Ravallion and Chen (2005) found that participants in a poor-area development program in China saved a large share of the income gains from the program. Identifying responses through savings and investment provides a clue to understanding current impacts on living standards and the possible future welfare gains beyond the project's current life span. Instead of focusing solely on the agreed welfare indicator relevant to the program's goals, one collects and analyzes data on a potentially wide range of intermediate indicators relevant to understanding the processes determining impacts.

This also illustrates a common concern in evaluation studies, given behavioral responses, namely that the study period is rarely much longer than the period of the program's disbursements. However, a share of the impact on peoples' living standards will usually occur beyond the disbursement period. This does not necessarily mean that credible evaluations will need to track welfare impacts over much longer periods than is typically the case—raising concerns about feasibility. But it does suggest that evaluations need to look carefully at impacts on partial intermediate indicators of longer-term impacts even when good measures of the welfare objective are available within the project cycle. The choice of such indicators will need to be informed by an understanding of participants' behavioral responses to the program. That understanding will be informed by economic theory and data.

In learning from an evaluation, one often needs to draw on information external to the evaluation. Qualitative research (intensive interviews with participants and administrators) can be a useful source of information on the underlying processes determining outcomes.⁹ One approach is to use such methods to test the assumptions made by an intervention; this has been called “theory-based evaluation,” although that is hardly an ideal term given that non-experimental identification strategies for mean impacts are often theory-based. Weiss (2001) illustrates this approach in the abstract in the context of evaluating the impacts of community-based anti-poverty programs. An example is found in a World Bank evaluation of social funds (SFs), as summarized in Carvalho and White (2004). While the overall aim of a SF is typically to reduce poverty, the study was interested in seeing whether SFs worked as intended by their designers. For example, did local communities participate? Who participated? Was there “capture” of the SF by local elites (as some critics have argued)? Building on Weiss (2001), the evaluation identified a series of key hypothesized links connecting the intervention to outcomes and tested whether each one worked. For example, in one of the country studies, Rao and Ibanez (2005) tested the assumption that a SF works by local communities collectively proposing the sub-projects that they want; for a SF in Jamaica, the authors found that the process was often dominated by local elites.

⁹ See the discussion on “mixed methods” in Rao and Woolcock (2003).

In practice, it is very unlikely that all the relevant assumptions are testable (including alternative assumptions made by different theories that might yield similar impacts). Nor is it clear that the process determining the impact of a program can always be decomposed into a neat series of testable links within a unique causal chain; there may be more complex forms of interaction and simultaneity that do not lend themselves to this type of analysis. For these reasons, theory-based evaluation cannot be considered an alternative to assessing impacts on final outcomes by credible (experimental or non-experimental) methods, although it can still be a useful complement to such evaluations, to better understanding measured impacts.

Project monitoring data bases are an important, under-utilized, source of information for understanding how a program works. Too often, however, the project monitoring data and the information system have negligible evaluative content. This is not inevitably the case. For example, Ravallion's (2000) method of combining spending maps with poverty maps can allow rapid assessments of the targeting performance of a decentralized anti-poverty program. This illustrates how, at modest cost, standard monitoring data can be made more useful for providing information on how the program is working and in a way that provides sufficiently rapid feedback to a project to allow corrections along the way.

The *Proempleo* experiment in Argentina provides an example of how information external to the evaluation can carry important insights into external validity. *Proempleo* was a pilot wage subsidy and training program for unemployed workers. The program's evaluation by Galasso et al. (2004) randomly assigned vouchers for a wage subsidy across (typically poor) people currently in a workfare program and tracked their subsequent success in getting regular work. A randomized control group located the counterfactual. The results did indicate a significant impact of the wage-subsidy voucher on employment. But when cross-checks were made against central administrative data, supplemented by informal interviews with the hiring firms, it was found that there was very low take-up of the wage subsidy by firms. The scheme was highly cost effective; the government saved 5% of its workfare wage bill for an outlay on subsidies that represented only 10% of that saving. However, the cross-checks against other data revealed that *Proempleo* did not work the way its design had intended. The bulk of the gain in employment for participants was not through higher demand for their labor induced by the wage subsidy. Rather the impact arose from supply side effects; the voucher had credential value to workers – it acted like a “letter of introduction” that few people had (and how it was allocated was a secret locally). This could not be revealed by the evaluation, but required supplementary data. The extra insight obtained about how *Proempleo* actually worked in the context of its trial setting also carried implications for scaling up, which put emphasis on providing better information for poor workers about how to get a job rather than providing wage subsidies.

Spillover effects also point to the importance of a deeper understanding of how a program operates. Indirect (or “second-round”) impacts on non-participants are common. A workfare program may lead to higher earnings for non-participants. Or a road improvement project in one area might improve accessibility elsewhere. Depending on how important these indirect effects are thought to be in the specific application, the

“program” may need to be redefined to embrace the spillover effects. Or one might need to combine the type of evaluation discussed here with other tools, such as a model of the labor market to pick up other benefits.

An extreme form of a spillover effect is an economy-wide program. The classic evaluation tools for assigned programs have little obvious role for economy-wide programs in which no explicit assignment process is evident, or if it is, the spillover effects are likely to be pervasive. When some countries get the economy-wide program but some do not, cross-country comparative work (such as growth regressions) can reveal impacts. That identification task is often difficult, because there are typically latent factors at country level that simultaneously influence outcomes and whether a country adopts the policy in question. And even when the identification strategy is accepted, carrying the generalized lessons from cross-country regressions to inform policy-making in any one country can be highly problematic. There are also a number of promising examples of how simulation tools for economy wide policies such as Computable General Equilibrium models can be combined with household-level survey data to assess impacts on poverty and inequality.¹⁰ These simulation methods make it far easier to attribute impacts to the policy change, although this advantage comes at the cost of the need to make many more assumptions about how the economy works.

In both assessing impacts and understanding the reasons for those impacts, there appears to be scope for a “meso” level analysis in which theory is used to inform empirical analysis of what would appear to be the key mechanisms linking an intervention to its outcomes, and this is done in a way that identifies key structural parameters that can be taken as fixed when estimating counterfactual outcomes. This type of approach can provide deeper insights into the factors determining outcomes in *ex post* evaluations and can also help in simulating the likely impacts of changes in program or policy design *ex ante*.

Naturally, simulations require many more assumptions about how an economy works.¹¹ As far as possible one would like to see those assumptions anchored to past knowledge built up from rigorous *ex post* evaluations. For example, by combining a randomized evaluation design with a structural model of education choices and exploiting the randomized design for identification, one can greatly expand the set of policy-relevant questions about the design of a program such as *PROGRESA* that a conventional evaluation can answer (Todd and Wolpin, 2002; Attanasio et al., 2004, and de Janvry and Sadoulet, 2006). This strand of the literature has revealed that a budget-neutral switch of the enrolment subsidy from primary to secondary school would have delivered a net gain in school attainments, by increasing the proportion of children who continue onto secondary school. While *PROGRESA* had an impact on schooling, it could have had a larger impact. However, it should be recalled that this type of program has two objectives: increasing schooling (reducing future poverty) and reducing current poverty, through the targeted transfers. To the extent that refocusing the subsidies on secondary schooling would reduce the impact on current income poverty (by increasing the forgone

¹⁰ See, for example, Bourguignon et al. (2003) and Chen and Ravallion (2004).

¹¹ For a useful overview of *ex ante* methods see Bourguignon and Ferreira (2003).

income from children's employment), the case for this change in the program's design would need further analysis.

These observations point to the important role played by theory in understanding why a program may or may not have impact. However, the theoretical models found in the evaluation literature are not always the most relevant to developing country settings. The models have stemmed mainly from the literature on evaluating training and other programs in developed countries, in which selection is seen largely as a matter of individual choice amongst those eligible. This approach does not sit easily with what we know about many anti-poverty programs in developing countries, in which the choices made by politicians and administrators appear to be at least as important to the selection process as the choices made by those eligible to participate. We often need a richer theoretical characterization of the selection problem to assure relevance. An example of one effort in this direction can be found in the Galasso and Ravallion (2005) model of a decentralized anti-poverty program; their model focuses on the public-choice problem facing the central government and the local collective action problem facing communities, with individual participation choices treated as a trivial sub-problem. Such models can also point to instrumental variables for identifying impacts and studying their heterogeneity.

An example of the use of a more structural approach to assessing an economy-wide reform can be found in Ravallion and van de Walle (2008). Here the policy being studied was the de-collectivization of agriculture in Vietnam, and subsequent efforts to develop a private market in land-use rights. These were huge reforms, affecting the livelihoods of the vast majority of the Vietnamese people. Ravallion and van de Walle develop a set of economic and political economy models that aim to explain how farmland was allocated to individual farmers at the time of de-collectivization, how those allocations affected living standards, and how the subsequent re-allocations of land amongst farmers (that were permitted by the subsequent agrarian reforms) responded to the inefficiencies left by the initial administrative assignment of land at the time of de-collectivization. Naturally, many more assumptions need to be made about how the economy works—essentially to make up for the fact that one cannot observe non-participants in these reforms as a clue to the counterfactual. Not all of the assumptions needed are testable. However, the principle of evaluation is the same, namely to infer the impacts of these reforms relative to explicit counterfactuals. For example, Ravallion and van de Walle assess the welfare impacts of the privatization of land-use rights against both an efficiency counterfactual (the simulated competitive market allocation) and an equity counterfactual (an equal allocation of quality-adjusted land within communes). This type of approach can also throw light on the heterogeneity of the welfare impacts of large reforms; in the Vietnam case, the authors were able to assess both the overall impacts on poverty (which were positive) and identify the presence of both losers and gainers, including among the poor.

VI. Publication Biases

The benefits from evaluative research depend on its publication. Development policy-making draws on accumulated knowledge built up in large part from published research findings. In addition to the contribution to shared knowledge through proper documentation and dissemination, publication has other benefits. It is an important screening and disciplining device for researchers; publishing in refereed professional journals helps establish a researcher's credibility (although one should never assume that publishing in even the most "prestigious" journals is a perfect indicator of research quality, given the mistakes made in editorial processes). Publishing also helps an institution's research department attract and keep the best researchers. Thus publication processes—notably the incentives facing journal editors and reviewers, researchers, and those who fund research—are relevant to our success in achieving development goals.

There are reasons for questioning how well the research publication process performs in helping to realize the social benefits from research. Three issues stand out. First, the cost of completing the publication stage in the cycle of research can be significant and it is hard to reduce these costs; writing the paper the right way, documenting everything that was done, addressing the concerns of referees and editors, all take time. Practitioners are often unwilling to fund these costs, and even question the need for publication. Again a large share of the benefits is external, to which individual practitioners naturally attach low weight.

Second, received wisdom develops its own inertia, through the publication process, with the result that it is often harder to publish a paper that reports unexpected or ambiguous impacts, when judged against current theories and/or past evidence. Reviewers and editors almost certainly apply different standards in according to whether they believe the results on *a priori* grounds. In the context of evaluating development projects, the prior belief is that the project will have positive impacts, for that is presumably the main reason why the project was funded in the first place. Then a bias toward confirming prior beliefs will mean that our knowledge is biased in favor of finding positive impacts. Negative or non-impacts will not get reported as easily. When there is a history of research on a type of intervention, the results of the early studies will set the priors against which later work is judged. An initial bad draw from the true distribution of impacts may then distort knowledge for some time after.

A third source of bias is that the review process in scientific publishing (at least in economics) tends to put greater emphasis on the internal validity of an evaluative research paper than on its external validity. The bulk of the effort goes into establishing that valid inferences are being drawn about causal impacts within the sample of program participants. The authors may offer some concluding (and possibly highly cautious) thoughts on the broader implications for scaling up the program well beyond that sample. However, these claims will rarely be established with comparable rigor to the efforts put into establishing internal validity, and the claims will rarely be challenged by reviewers.

These imperfections in the research publication industry undoubtedly have feedback effects on the production of evaluative research. Researchers will tend to work harder to obtain positive findings, or at least results consistent with received wisdom, so as to improve their chances of getting their work published. No doubt, extreme biases (in either direction) will be eventually exposed. But this takes time.

Researchers have no shortage of instruments at their disposal to respond to the (often distorted) incentives generated by professional publication processes. Key decisions on what to report, and indeed the topic of the research paper, naturally lie with the individual researcher, who must write the paper and get it published. In the case of impact evaluations of development projects, the survey data (often collected for the purpose of the evaluation) will typically include multiple indicators of “outcomes.” If one collects 20 indicators (say) then there is a good chance that at least one of them shows statistically significant impacts of the project even when it had no impact in reality. A researcher keen to get published might be tempted to report results solely for the significant indicator. (Journal reviewers and editors rarely ask what other data were collected.) The dangers to knowledge generation are plain.

The threat of replication by another researcher can help assure better behavior. But in economics, replication studies tend to have low status and are actually quite rare. Nor do researchers have a strong incentive to make their data publicly available for replication purposes. Some professional economics journals have adopted a policy that data sets should be made available this way, although enforcement does not appear to be strong.

In choosing how to respond to this environment, individual researchers face a trade-off between publishability and relevance. Thankfully, the fact of being policy relevant is not in itself an impediment to publishability in most journals, though any research paper that lacks originality, rigor or depth will have a hard time getting published. It is by maintaining the highest standards that we assure that relevant research is publishable, as well as being credible when carried to policy dialogues. However, it must be acknowledged that the set of research questions that are most relevant to development policy overlap only partially with the set of questions that are seen to be in vogue by the editors of the professional journals, at any given time. The dominance of academia in the respected publishing outlets is understandable, but it can sometimes make it harder for researchers doing work more relevant to development practitioners, even when that work meets the standards of more academic research. Academic research draws its motivation from academic concerns that overlap imperfectly with the issues that matter to development practitioners. Provided that scholarly rigor is maintained, the cost to a researcher’s published output of doing policy relevant research might not be high, but it would be naïve to think that the cost is zero.

Communication and dissemination of the published findings from evaluative research can also be deficient. Researchers sometimes lack the skills or personalities needed for effective communication with non-technical audiences. Having worked very hard to assure that the data and analysis are sound, and so pass muster by accepted scientific criteria, it does not come easily for all researchers to translate the results into

just a few key policy messages, which do not seem to do justice to all the work involved. The externality problem can also arise here, whereby social returns from outreach exceed private returns. A research institution will often need to support its researchers with specialized staff, with strong communication skills.

Conclusions

Knowledge externalities entail that we probably under-invest in evaluative research. This problem appears to be particularly severe for certain types of research activities, notably evaluations of development interventions that yield benefits over long periods and for efforts in the production of relevant data and other tools for improving development practice. The process of knowledge generation through evaluations is probably also affected by biases on the publication side, which distort the incentives facing individual researchers in producing evaluative research.

None of this is helped by the fact that the favored approaches of evaluators often fall well short of delivering credible answers to the questions posed by practitioners. On the one hand, exaggerated claims are sometimes made about what can be learnt about development effectiveness in a short time with little or no credible data. On the other hand, the more inferentially sound (and costly) methods have focused on a rather narrow subset of the questions relevant to practitioners.

Effort is needed to develop approaches to evaluation that can throw more useful light on the external validity of findings on specific projects (including implications for scaling up) and can provide a deeper understanding of what determines why an intervention does, or does not, have impact. There is still much to do if we want to realize the potential for evaluative research to inform development policy by “seeking truth from facts.”

References

- Attanasio, Orazio, Costas Meghir and Miguel Szekely, 2003, "Using Randomized Experiments and Structural Models for Scaling Up: Evidence from the PROGRESA Evaluation," Working Paper EWP04/03, Institute of Fiscal Studies London.
- Attanasio, Orazio, Costas Meghir and Ana Santiago, 2004, "Education Choices in Mexico: Using a Structural Model and a Randomized Experiment to Evaluate PROGRESA," Working Paper EWP04/04, Institute of Fiscal Studies London.
- Banerjee, Abhijit, 2007, *Making Aid Work*, Cambridge, Mass.: MIT Press.
- Behrman, Jere, Yingmei Cheng and Petra Todd, 2004, "Evaluating Preschool Programs When Length of Exposure to the Program Varies: A Nonparametric Approach," *Review of Economics and Statistics*, 86(1): 108-32.
- Björklund, Anders and Robert Moffitt, 1987, "The Estimation of Wage Gains and Welfare Gains in Self-Selection," *Review of Economics and Statistics* 69(1): 42-49.
- Bourguignon, François and Francisco Ferreira, 2003, "Ex-ante Evaluation of Policy Reforms Using Behavioural Models," in Bourguignon, F. and L. Pereira da Silva (eds.) *The Impact of Economic Policies on Poverty and Income Distribution*, New York: Oxford University Press.
- Carvalho, Soniya and Howard White, 2004, "Theory-Based Evaluation: The Case of Social Funds," *American Journal of Evaluation* 25(2): 141-160.
- Chen, Shaohua, Ren Mu and Martin Ravallion, 2007, "Are There Lasting Impacts of Aid to Poor Areas? Evidence from Rural China," Policy Research Working Paper 4084, World Bank, Washington DC.
- Chen, Shaohua and Martin Ravallion, 2007, "Absolute Poverty Measures for the Developing World," *Proceedings of the National Academy of Sciences of the United States of America*, 104(43): 16757-16762.
- Deaton, Angus, 2006, "Evidence-based aid must not become the latest in a long string of development fads," *Boston Review* July; <http://bostonreview.net/BR31.4/deaton.html>
- De Janvry, Alain and Elisabeth Sadoulet, 2006, "Making Conditional Cash Transfer Programs More Efficient: Designing for Maximum Effect of the Conditionality," *World Bank Economic Review* 20(1): 1-29.
- Demirguc-Kunt, Asli and Enrica Detragiache, 2002, "Does Deposit Insurance Increase Banking System Stability? An Empirical Investigation," *Journal of Monetary Economics* 49(7): 1373-1406.

- Djebbari, Habiba and Jeffrey Smith, 2005, "Heterogeneous Program Impacts of PROGRESA," mimeo, Laval University and University of Michigan.
- Du Runsheng, 2006, *The Course of China's Rural Reform*, International Food Policy Research Institute, Washington DC.
- Duflo, Esther and Michael Kremer, 2005, "Use of Randomization in the Evaluation of Development Effectiveness," in George Pitman, Osvaldo Feinstein and Gregory Ingram (eds.) *Evaluating Development Effectiveness*, New Brunswick, NJ: Transaction Publishers.
- Feder, Gershon, T. Onchan, Y. Chalamwong and C. Hongladarom, 1988, *Land Policies and Farm Productivity in Thailand*, Johns Hopkins University Press.
- Galasso, Emanuela and Martin Ravallion, 2004, "Social Protection in a Crisis: Argentina's *Plan Jefes y Jefas*," *World Bank Economic Review*, 18(3): 367-399.
- _____ and _____, 2005, "Decentralized Targeting of an Anti-Poverty Program," *Journal of Public Economics*, 85: 705-727.
- Galasso, Emanuela, Martin Ravallion and Agustin Salvia, 2004, "Assisting the Transition from Workfare to Work: Argentina's *Proempleo* Experiment", *Industrial and Labor Relations Review*, 57(5): 128-142.
- Heckman James and Jeffrey Smith, 1995, "Assessing the Case for Social Experiments," *Journal of Economic Perspectives* 9(2): 85-110.
- Heckman, James, Hidehiko Ichimura, Jeffrey Smith, and Petra Todd, 1998, "Characterizing Selection Bias using Experimental Data," *Econometrica* 66, 1017-1099.
- Heckman, James, Jeffrey Smith and Nancy Clements, 1997, "Making the Most Out of Programme Evaluations and Social Experiments: Accounting for Heterogeneity in Programme Impacts," *Review of Economic Studies* 64(4), 487-535.
- Heckman, James, L. Lochner and C. Taber, 1998, "General Equilibrium Treatment Effects," *American Economic Review Papers and Proceedings* 88: 381-386.
- Heckman, James and Salvador, Navarro-Lozano, 2004, "Using Matching, Instrumental Variables and Control Functions to Estimate Economic Choice Models," *Review of Economics and Statistics* 86(1): 30-57.
- Heckman, James, Serio Urzua and Edward Vytlacil, 2006, "Understanding Instrumental Variables in Models with Essential Heterogeneity," *Review of Economics and Statistics* 88(3): 389-432.
- Hertel, Thomas and L. Alan Winters, 2006, *Poverty and the WTO: Impacts of the Doha Development Agenda*, New York: Palgrave Macmillan.

- Hoekman, Bernard and Marcelo Olarreaga (eds), 2007, *Global Trade and Poor Nations: The Poverty Impacts and Policy Implications of Liberalization*, Brookings Institution, Washington DC.
- Lanjouw, Peter and Martin Ravallion, 1999, "Benefit Incidence and the Timing of Program Capture," *World Bank Economic Review* 13(2): 257-274.
- Luo, Xiaopeng, 2007, "Collective Learning Capacity and the Choice of Reform Path." Paper presented at the IFPRI/Government of China Conference, "Taking Action for the World's Poor and Hungry People," Beijing.
- Miguel, Edward and Michael Kremer, 2004, "Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities," *Econometrica* 72(1): 159-217.
- Moffitt, Robert, 2003, "The Role of Randomized Field Trials in Social Science Research: A Perspective from Evaluations of Reforms of Social Welfare Programs," Cemmap Working Paper, CWP23/02, Department of Economics, University College London.
- _____, 2004, "The Role of Randomized Field Trials in Social Science Research," *American Behavioral Scientist* 47(5): 506-540.
- _____, 2006, "Forecasting the Effects of Scaling Up Social Programs: An Economics Perspective," in Barbara Schneider and Sarah-Kathryn McDonald (eds) *Scale-Up in Education: Ideas in Principle*. Rowman and Littlefield.
- Organization for Economic Co-Operation and Development, 2007, *A Practical Guide to Ex Ante Poverty Impact Assessment*, Development Assistance Committee Guidelines and Reference Series, OECD, Paris.
- Rao, Vijayendra and Ana Maria Ibanez, 2005, "The Social Impact of Social Funds in Jamaica: A Mixed Methods Analysis of Participation, Targeting and Collective Action in Community Driven Development," *Journal of Development Studies* 41(5): 788-838.
- Rao Vijayendra and Michael Walton (eds), 2004, *Culture and Public Action*, Stanford: Stanford University Press.
- Rao, Vijayendra and Michael Woolcock, 2003, "Integrating Qualitative and Quantitative Approaches in Program Evaluation," in Francois J. Bourguignon and Luiz Pereira da Silva (eds.) *The Impact of Economic Policies on Poverty and Income Distribution: Evaluation Techniques and Tools* New York: Oxford University Press, pp. 165-90.
- Ravallion, Martin, 2000, "Monitoring Targeting Performance when Decentralized Allocations to the Poor are Unobserved," *World Bank Economic Review* 14(2): 331-45.

- _____, 2004, "Who is Protected from Budget Cuts?" *Journal of Policy Reform*, 7(2): 109-22.
- _____, 2006, "Looking Beyond Averages in the Trade and Poverty Debate," *World Development* (special issue on *The Impact of Globalization on the World's Poor*, edited by Machiko Nissanke and Erik Thorbecke), 34(8): 1374-1392.
- _____, 2007, "How Relevant is Targeting to the Success of an Antipoverty Program?" Policy Research Working Paper 4358, World Bank, Washington DC.
- _____, 2008, "Evaluating Anti-Poverty Programs," in *Handbook of Development Economics Volume 4*, edited by Paul Schultz and John Strauss, Amsterdam: North-Holland.
- Ravallion, Martin and Shaohua Chen, 2005, "Hidden Impact: Household Saving in Response to a Poor-Area Development Project," *Journal of Public Economics*, 89: 2183-2204.
- _____ and _____, 2007, "China's (Uneven) Progress Against Poverty," *Journal of Development Economics*, 82(1): 1-42.
- Ravallion, Martin and Gaurav Datt, 1995. "Is Targeting through a Work Requirement Efficient? Some Evidence for Rural India," in D. van de Walle and K. Nead (eds) *Public Spending and the Poor: Theory and Evidence*, Baltimore: Johns Hopkins University Press.
- Ravallion, Martin, Gaurav Datt and Dominique van de Walle, 1991, "Quantifying Absolute Poverty in the Developing World," *Review of Income and Wealth*, 37: 345-361.
- Ravallion, Martin, Emanuela Galasso, Teodoro Lazo and Ernesto Philipp, 2005, "What Can Ex-Participants Reveal About a Program's Impact?" *Journal of Human Resources*, 40 (Winter): 208-230.
- Ravallion, Martin, Dominique van de Walle and Madhur Gaurtam, 1995, "Testing a Social Safety Net," *Journal of Public Economics*, 57(2): 175-199.
- Ravallion, Martin and Dominique van de Walle, 2008, *Land in Transition: Reform and Poverty in Rural Vietnam*, Palgrave Macmillan, forthcoming.
- Sadoulet, Elisabeth, Alain de Janvry and Benjamin Davis, 2001, "Cash Transfer Programs with Income Multipliers: PROCAMPO in Mexico," *World Development* 29(6): 1043-56.
- Todd, Petra and Kenneth Wolpin, 2002, "Using a Social Experiment to Validate a Dynamic Behavioral Model of Child Schooling and Fertility: Assessing the Impact of a School Subsidy Program in Mexico," Penn Institute for Economic

Research Working Paper 03-022, Department of Economics, University of Pennsylvania.

_____ and _____, 2006, “Ex-Ante Evaluation of Social Programs,” mimeo, Department of Economics, University of Pennsylvania.

van de Walle, Dominique and Ren Mu, 2007, “Fungibility and the Flypaper Effect of Project Aid: Micro-Evidence for Vietnam,” *Journal of Development Economics* 84: 667-685.

Weiss, Carol, 2001, “Theory-Based Evaluation: Theories of Change for Poverty Reduction Programs,” in O. Feinstein and R. Piccioto (eds), *Evaluation and Poverty Reduction*, New Brunswick, NJ: Transaction Publications.

World Bank, 1990, *World Development Report: Poverty*, New York: Oxford University Press.