Evaluation in the Practice of Development

Martin Ravallion

Standard methods of impact evaluation often leave significant gaps between what we know about development effectiveness and what we want to know—gaps that stem from distortions in the market for knowledge. The author discusses how evaluations might better address these knowledge gaps and so be more relevant to the needs of practitioners. It is argued that more attention needs to be given to identifying policy-relevant questions (including the case for intervention), that a broader approach should be taken to the problems of internal validity (including heterogeneity and spillover effects), and that the problems of external validity (including scaling up) merit more attention by researchers. JEL codes: H43, O22

Anyone who doubts the potential benefits to development practitioners from evaluation should study China's economic reforms. In 1978, the Communist Party's 11th Congress broke with its ideology-based view of policymaking in favor of a pragmatic approach, which Deng Xiaoping famously dubbed "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—"the intellectual approach of seeking truth from facts" (Du 2006, p. 2). In looking for facts, a high weight was put on demonstrable success in actual policy experiments on the ground. The first major application was to rural reform. While there had been much dissatisfaction with collectivized farming, there were competing ideas as to what needed to be done. The evidence from local experiments was eventually instrumental in persuading even the old guard of the Party's leadership (many of whom still favored collectivized farming) that household contracts could deliver higher food output. The evidence had to be credible. A new research group did field work studying the local experiments—though they were certainly not randomized experiments—in using contracts with individual farmers. The evidence might not be

doi;10.1093/wbro/lkp002 Advance Access publication March 25, 2009 24:29–53

conclusive by today's scientific standards, but it helped to convince skeptical policymakers (many still imbued in Maoist ideology) of the merits of scaling up the local initiatives (Luo 2007). The rural reforms that were then implemented nationally helped achieve what was 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 policymaking elsewhere. (And China still has much that it could do to enhance the credibility of its own efforts at evidence-based policymaking.) In this paper I argue that we underinvest in rigorous evaluations of development interventions *and* that the evaluations that are done currently are not as useful as they could be. Distortions in the "market for knowledge" about development effectiveness leave persistent gaps between what we know and what we want to know; and the learning process is often too weak to guide practice reliably. The outcome is almost certainly one of less overall impact on poverty.

I first try to understand how the gaps in our knowledge about development effectiveness come to exist and persist. I then identify a number of things that need to change in current approaches to evaluation if the potential to inform development practice is to be fulfilled. Examples are given from recent research, although a number of these issues remain under-researched. I hope that the discussion will help to change that.

Why Might We Underinvest in Rigorous Evaluations?

The focus of this paper is the problem of assessing the impact of a development project, where "impact" is measured against explicit counterfactual outcomes (such as in the absence of the project); the essential characteristic of a rigorous evaluation is that it includes a credible strategy for identifying the counterfactual. The topic embraces both *ex ante* and *ex post* evaluation (and possibly both for the same project). *Ex ante* evaluation is a key input to project appraisal. *Ex post* evaluation can sometimes provide useful insights into how a project might be modified along the way, and is certainly a key input to the accumulation of knowledge about development effectiveness, which guides future policymaking.

There are good reasons why not everything that is done in the name of development gets evaluated. Rigorous evaluations are rarely easy. Practical and logistical difficulties abound. 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.

However, there are also reasons to doubt that the market for knowledge about development effectiveness works well. The outcome of these market failures is almost certainly that we underinvest in rigorous impact evaluations.

Suppliers and demanders of knowledge about development effectiveness do not typically have the same information about the quality of the evaluation—giving an example of what economists call "asymmetric information," which is a well-known source of market failure.[1] In the present context, development practitioners cannot easily assess the quality and expected benefits of an impact evaluation in order to weigh them against the costs. Short-cut methods promise quick results at low cost, though rarely are users well informed of the inferential dangers.[2] Since it is often hard for practitioners to know whether research is of good quality or not, there is a real risk that rigorous evaluations are driven out by nonrigorous ones of doubtful veracity.

Another important feature of this market is the degree of control that individual project "managers" (including staff in both aid agencies and governments) have over what gets evaluated and how much is spent on evaluation. This can be thought of as a noncompetitive feature of the market for knowledge about development effectiveness, in that the project manager more or less has the power to block the supply of knowledge. 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, or by political stakeholders whose incentives need not be well-aligned with knowledge demands. The portfolio of evaluations is almost certainly biased towards programs that work well; managers of weak programs try to avoid rigorous evaluation, which threatens to expose the program's weaknesses. Lighter "evaluations" are often easier to manipulate for the purpose of showing seemingly positive results.

Decentralized decision-making about evaluation generates another source of market failure: the benefits from the rigorous evaluation of a development project spill over to other projects, which typically do not share in the cost of doing that evaluation. 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 evaluation. This is what economists call an "externality." An implication of the externalities in the market for knowledge about development effectiveness is that we tend to underinvest in research that can draw useful lessons for other projects and settings besides that of the specific evaluation.

Certain types of evaluations are likely to be more prone to these sources of market failure. It is typically far easier to evaluate an intervention that yields all its likely impacts within 1 year (say) 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 transfer payments 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 often 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 underinvest in evaluations of types of interventions that tend to have diffused, widespread 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 nonparticipants (as the basis for inferring the counterfactual). It may also be hard to fund evaluations for such interventions, since they often lack a well-defined constituency of political support.

The implication of all this is that, without strong institutional support and encouragement, there will probably be too few evaluations, particularly of the long-term impacts of development interventions and of broader sectoral or economy-wide reforms. And the evaluations that do get done will focus too much on *internal validity* (whether valid inferences are drawn about the impact of that specific project in its specific setting) relative to *external validity* (whether valid inferences are drawn for other projects, either as scaled up versions of that project in the same setting or as similar projects in different settings). The fact that long-term evaluations are so rare (though it is widely agreed that development does not happen rapidly) and that we clearly know too little about external validity suggest that the available support is currently insufficient *or* it is misallocated.

Rising Support for Evaluations

Increasingly evaluations do 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 increasingly being pressed by their citizens to show the impact of development aid which has generated extra resources for financing impact evaluations. Unfortunately, the resources available are not always used for making rigorous evaluations. And it is not clear that the extra resources are having as much impact as they could on the incentives facing project managers and governments. Donor support needs to focus on increasing marginal private benefits from evaluation or reducing marginal costs. Nonetheless, there is now a 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.

What does get evaluated, however, is still only a modest fraction of what gets done on the ground in the name of development. That may always be the case, given the costs of evaluation. But what is more worrying is that this fraction is a decidedly *nonrandom* one. Typically, 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.

All this makes it very important that new efforts by the development community to support impact evaluations of development policies—to address the market failures discussed above—should start from those knowledge gaps, not from a researcher's prior preference for one sort of data or method. That is not always the case. For example, while the recent enthusiasm for Randomized Control Trials (RCTs) (also called social experiments)—see, for example, Banerjee (2007) and Duflo and Kremer (2005)—has generated some interesting new research, it is not based on any clear strategic assessment of how this particular method would fill the knowledge gaps of highest priority. Nor is there any obvious reason why doing more social experiments would help correct for the distortions that generated those knowledge gaps. Randomization is clearly only feasible for a nonrandom subset of policies and settings; for example, it is rarely feasible to randomize the location of infrastructure projects and related programs, which are core activities in almost any poor country's development strategy. And even for the types of programs for which randomization is an option, it will be adopted more readily in some settings than others, given that social experiments raise ethical and political concerns—stemming from the fact that some of those to which a program is randomly assigned will almost certainly not need it, while some in the control group will. A better idea would be to randomize what gets evaluated rigorously and then choose a method appropriate to each sampled intervention, with randomization as one option.

The rest of this article will explore how we might assure that future work on impact evaluation is more relevant to the needs of development practitioners. While better approaches to evaluation will not, on their own, solve all the problems in the market for knowledge discussed above, recognizing those problems is the logical starting point for thinking about what constitutes better evaluation.

How Can We Do Better in Filling Key Knowledge Gaps?

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 (nonparticipants). While this is an undeniably important and challenging problem, solving it is not sufficient for assuring that evaluation is relevant to development practice.

Questions for Evaluations

Evaluations should not take the intervention as predetermined, 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? As Devarajan and others (1997) argue, 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 subgroups 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 in conceptualizing the case for intervention can be especially important when the capacity for development policymaking 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 prior beliefs about what needs to be done. Over time, some practitioners become experts at specific types of intervention, 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. Too often it is not the question that is driving the evaluation 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 around. Starting with the question, not the method, often points the evaluator toward types of data and methods outside the domain traditionally favored by his or her own disciplinary background. For example, some of the World Bank's research economists trying to understand persistent poverty and the impacts of antipoverty programs have been drawn to 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 already noted, rigorous impact evaluations require credible strategies for identifying the counterfactual—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 nonparticipants are compared at only one date; see Ravallion (2008) for a survey of the (experimental and nonexperimental) methods available for this task. This is all about internal validity, which has been the main focus of researchers working on evaluations. In this discussion I will flag some issues that have received less attention yet matter greatly to the impact of an evaluation.

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 policymakers. The alternative of doing nothing is rarely of interest to policymakers, who prefer instead 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 it is still performing 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 (untargeted) 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 an interest in better understanding the *horizontal impacts* of a program, that is 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 standard impact 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 and others (1995) for the purpose of measuring the impacts of changes in social spending on the intertemporal 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). (This is only possible if

one can identify how impacts vary with household characteristics; the discussion will return to this issue in discussing impact heterogeneity below.) Ravallion and others 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 spill over to nonparticipants. Only under this assumption can we infer the counterfactual from an appropriate sample of the nonparticipants. 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; spillover are also relevant in drawing lessons for scaling up based on an RCT. For further discussion, see Moffitt (2003, 2006).

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 an evaluation 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. Consider 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; 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 "nonparticipants." 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 spill over 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 nonparticipants. 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 and others (2009) 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 nonprogram villages—the same set of villages from which the comparison group was drawn. Ignoring these spillover effects generated a nonnegligible underestimation of the impact of the program. Chen and others show how, under certain assumptions, 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. In the case of the poor-area program in China that Chen and others study, their results suggest that the spending responses of local governments to the external aid entail that the standard "difference-in-difference" method may well capture only two-thirds of the true impact.

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 to 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.

By recognizing the scope for heterogeneity in impacts and the role of contextual factors, one can make evaluative research more relevant to good policymaking. For example, in the aforementioned evaluation of a poor-area development program in rural China, Chen and others (2009) find low overall impact but considerable heterogeneity, in that different types of households benefited more than others, with the relatively better educated amongst the poor achieving the highest returns to the project's investments. The policy implication is that choosing different beneficiaries would have greatly increased the project's overall impact; indeed, the study estimated that an alternative process of beneficiary selection that better exploited the heterogeneity in impacts could have led to a four-fold increase in the

project's overall rate of return. By developing a deeper understanding of such heterogeneity, evaluations can help develop better projects.

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. However, 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 experiments; this is what Heckman and others (2006) refer to as "essential heterogeneity." 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 approach to the evaluation data for PROGRESA (a conditional cash transfer program in Mexico), Djebbari and Smith (2008) found that they could convincingly reject the assumption of common (homogeneous) effects made by past evaluations of that program.

When there is such heterogeneity, it can be of interest to policymakers to distinguish marginal impacts (from small program expansions or contractions) from the average impacts that have received the bulk of attention. 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 and others (1997).

External Validity

Arguably the most important thing to learn from any evaluation relates to its lessons for future policies (including reforms to the interventions being evaluated). External validity is highly desirable, but 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 *why* a policy or program has an impact; a question to which I will return. 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, or for taking the same project to other settings. Research that tests the theories that underlie the rationales for 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 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 and others (2005) of the impacts on workfare participants of leaving the program relative to staying (recognizing that this entails a nonrandom selection process). Another example can be found in the study by Behrman and others (2004) of the impacts on children's cognitive skills and health status of longer exposure to a preschool program in Bolivia. The authors provided 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, one must recognize the importance of *context* since this can be key to drawing valid lessons for other settings. Relevant contextual factors may include the circumstances of participants, the economic, cultural, and political environment, and the administrative context. Unless we understand how such factors influence the outcomes of an intervention, the evaluation will have weak external validity. The next section returns to this issue.

Given that we can expect in general that any intervention will have heterogeneous impacts—some participants gain more than others—serious concerns can arise about the external validity of RCTs. 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 were included in the trial.[3] The RCT may well have evaluated a very different program to the one that is actually implemented on the basis of that RCT.

External validity concerns about impact evaluations can also arise when certain institutions need to be presented 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 the national scale may have a very different impact in places where the NGO is not present. 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 (among other things) the quality of supervision, the incentives facing service providers, and administrative capacity.

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 may be deceptive when scaled up, given that the structure of returns to schooling will alter. Heckman and others (1998) demonstrated that partial equilibrium analysis can greatly overestimate the impact of a tuition subsidy once relative wages adjust, although Lee (2005) found 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) gave 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) discussed how this can happen in theory and presented the 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 underestimate 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 nonpoor (as shown by Lanjouw and Ravallion 1999, using data for India).

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 an 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 questionable 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, whereby the marginal use of public finds is unlikely to be the specific project that is being evaluated. Yet an assessment of "aid effectiveness" is (presumably) just that—an evaluation of the impact of the aid, not the project *per se*. These are different evaluation problems.

Assessments of aid effectiveness need to take a broader view of public spending, as advocated by Devarajan and others (1997). How broad it needs to be is unclear. There is some evidence that external aid sticks to its sector (quaintly

called a "flypaper effect" in economics); on this see van de Walle and Mu (2007). The existence of fungibility and flypaper effects points to the need for a sectoral approach in efforts to evaluate the impacts of development aid.

What Determines Impact?

The above discussion points to the need to supplement standard evaluations by information that can throw light on the factors influencing measured outcomes. That can be crucial for drawing useful policy lessons, including redesigning a program and scaling up. 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 how the role of local institutions influences outcomes). This section elaborates some of the ways that we might learn more about how a program does, or does not, have an impact, so as to better address the issues raised above.

An obvious approach to understanding which 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 RCTs across varying contexts and scales should be used to decide what works and what does not in development aid. Even putting aside the aforementioned problems encountered in 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 number of RCTs needed to test even one large national program could well be prohibitive. It is questionable whether this is a sound strategy for filling the existing gaps in our knowledge about development effectiveness.

Nonetheless, even if one cannot go as far as Banerjee (2007) would like, it can be agreed that evaluation designs should plan for contextual variation. Important clues 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 (2005) in which the targeting performance of Bangladesh's Food-for-Education program was assessed across each of 100 villages in Bangladesh, with the results being correlated with the 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 intravillage 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 United States.

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 and others (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.

Useful clues for understanding impacts can sometimes be found by studying impacts on what can be called "intermediate" or "structural" measures. The typical evaluation design identifies a small number of "final outcome" indicators, and it 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 deemed relevant on theoretical grounds. For example, the intertemporal behavioral responses of participants in antipoverty 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, de Janvry, and Davis 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 understand the 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 both 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; see the discussion on "mixed methods" in Rao and Woolcock (2003). 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 identification strategies for mean impacts are often theory based. Weiss (2001) illustrated this approach in the abstract in the context of evaluating the impacts of community-based antipoverty 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 an 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 an SF works by local communities collectively proposing the subprojects that they want; for an SF in Jamaica, the authors found that the process was often dominated by local elites.

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 nonexperimental) methods, although it can still be a useful complement to such evaluations for better understanding measured impacts.

Project monitoring databases are an important, underutilized, source of information for understanding how a program works. Too often, however, the project monitoring data collected and the information system used 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 antipoverty 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. Proempleo was a pilot wage subsidy and training program for unemployed workers. The RCT by Galasso and others (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 indicated 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 percent of its workfare wage bill for an outlay on subsidies that represented only 10 percent of that saving. However, the cross-checks against these 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 appeared to have 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 nonparticipants are common. A workfare program may lead to higher earnings for nonparticipants; 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 policymaking 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.[4] 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 is often 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.[5] 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 that a conventional evaluation can answer; examples using the PROGRESA evaluation data can be found in Todd and Wolpin (2002), Attanasio and others (2004), and de Janvry and Sadoulet (2006). This strand of the literature has revealed that a budget-neutral switch of the enrolment subsidy in PROGRESA 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 greater 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 income from children's employment), the case for this change in the program's design would need further analysis.

Many of these observations point to the important role played by theory in understanding why a program may or may not have an 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 antipoverty 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 antipoverty program; their model focuses on the public-choice problem facing the government and the local collective action problem facing communities, with individual participation choices treated as a trivial subproblem. 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 decollectivization of agriculture in Vietnam and the 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 developed models to explain how farmland was allocated to individual farmers at the time of decollectivization, how those allocations affected living standards, and how the subsequent reallocations of land amongst farmers (that were permitted by the subsequent market-oriented agrarian reforms) responded to the inefficiencies left by the initial administrative assignment of land at the time of decollectivization. Naturally, many more assumptions need to be made about how the economy works—essentially to make up for the fact that one cannot observe nonparticipants in these reforms as a clue to the counterfactual. Not all of those assumptions 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 assessed 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 and identify the presence of both losers and gainers, including among the poor.

Does Published Knowledge Reliably Guide Development Practice?

The benefits from evaluations depend in part on their publication, which is the main way they feed into development knowledge. Development policymaking draws on accumulated knowledge built up in large part from published findings.

At the same time, publishing in refereed journals is important to a researcher's credibility and career prospects. 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 publication process performs in helping to realize the social benefits from rigorous evaluations. 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 they even question the need for publication. Again a large share of the benefits is external, to which individual project staff 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, past evidence, or both. Reviewers and editors are likely to apply different standards according to whether they believe the results hold on *a priori* grounds.

In the context of evaluating development projects, the prior belief is often 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 preference for confirming prior beliefs will tend to bias our knowledge in favor of finding positive impacts. Negative or nonimpacts 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 prior beliefs 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 are rarely challenged by reviewers.

These imperfections in the research publication industry undoubtedly have feedback effects on the production of evaluations. 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. Thus, as Rodrik (2009) points out, there will be little or no incentive for researchers to carry out the great many repetitions that would probably be called for in the agenda for the mass RCTs proposed by Banerjee (2007) and Duflo and Kremer (2005), given that professional journals would have little interest in such replications of the same intervention and method in different settings.

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 the datasets used in accepted papers should be made available this way, although enforcement is not uniformly strong.

In choosing how to respond to this environment, the individual researcher faces 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 academic standards. 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 on development effectiveness can also be deficient. Researchers sometimes lack the skills or personalities needed for effective communication with nontechnical 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 who possess strong communication skills.

Conclusions

We underinvest in some of the most important tools for enhancing development effectiveness. Weak incentives facing key decision-makers—stemming from knowledge externalities, asymmetric information, and noncompetitive features of the market for knowledge—entail that too few rigorous impact evaluations of development interventions get done. This problem appears to be particularly severe for evaluations of projects that yield benefits over long periods and for efforts in rigorously understanding the lessons that can be drawn for other projects and settings. While donor support for evaluation is helping redress these problems, there is still a long way to go; greater support is needed, but existing support could also be made more effective if it were aimed at changing private incentives to evaluate, by either raising the marginal benefits or lowering the marginal costs facing project managers. 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 doing evaluations.

None of this is helped by the fact that even the most rigorous methods found in practice often fall well short of delivering credible answers to the questions posed by practitioners. Those questions start at the outset of the project cycle and even embrace the rationale for the intervention. They include understanding why the intervention might have greater impact for some participants, and in some settings, than others. They include the lessons for both the intervention under study and (importantly) future interventions. The classic estimate of the mean impact on those treated is of strictly limited utility for addressing these issues.

Nor is the task helped by the fact that researchers have at times overstated what their favorite method can deliver for practitioners, and that they have often chosen what they evaluate according to whether their favorite method is feasible, rather than whether the question is important to development. Interventions are even being chosen, or designed, to fit certain preferred evaluation methods. At the same time, exaggerated claims are sometimes made by nonresearchers about what can be learnt about development effectiveness in a short time with little or no credible data.

Looking forward, greater 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 that can provide a deeper understanding of what determines why an intervention does, or does not, have an impact. Fungibility and flypaper effects also point to the need for a broader sectoral approach to assessing aid effectiveness. 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."

Notes

Martin Ravallion is Director, Development Research Group, World Bank; his email address is Mravallion@worldbank.org. For helpful comments on an earlier version of this article, and related discussions on this topic, the author is grateful to Francois Bourguignon, Asli Demirguc-Kunt, Gershon Feder, Jed Friedman, Emanuela Galasso, Markus Goldstein, Bernard Hoekman, Beth King, Danny Leipziger, David McKenzie, Luis Seven, Lyn Squire, Dominique van de Walle, Michael Woolcock, and the journal's reviewers. These are the views of the author and should not be attributed to the World Bank or any affiliated organization.

1. The classic account of this problem is given in Akerlof (1970).
2. For example, OECD (2007) outlines an approach to "*ex ante* poverty impact assessment" that claims to assess the "poverty outcomes and impacts" of a project in just 2–3 weeks at a cost of $10,000–40,000, which, as the authors point out, is appreciably less than standard impact evaluations. The OECD paper proposes that a consultant fills in a series of tables giving the project's "short-term and long-term outcomes" across a range of (economic and noneconomic) dimensions for each of the various groups of identified "stakeholders," as well as the project's "transmission channels," through induced changes in prices, employment, transfers, and so on. Many readers (including many practitioners) would not know just how hard it is to make such assessments in a credible way, and the paper offers no guidance to readers on what degree of confidence one can have in the results of such an exercise.
3. This is sometimes called "randomization bias"; see Heckman and Smith (1995). See also the discussion in Moffitt (2004).
4. See, for example, Bourguignon and Ferreira (2003) and Chen and Ravallion (2004).
5. For a useful overview of *ex ante* methods, see Bourguignon and Ferreira (2003).

References

Akerlof, George. 1970. "The Market for 'Lemons': Quality Uncertainty and the Market Mechanism." *Quarterly Journal of Economics* 84:488–500.

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. London: Institute of Fiscal Studies.

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. London: Institute of Fiscal Studies.

Banerjee, Abhijit. 2007. *Making Aid Work*. Cambridge, MA: 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):42–9.

Bourguignon, François, and Francisco Ferreira. 2003. "Ex-ante Evaluation of Policy Reforms Using Behavioural Models." In Francois F Bourguignon, and Luiz 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):141–60.

Chen, Shaohua, and Martin Ravallion. 2004. "Welfare Impacts of China's Accession to the World Trade Organization." *World Bank Economic Review* 18(1):29–58.

Chen, Shaohua, Ren Mu, and Martin Ravallion. 2009. "Are there Lasting Impacts of Aid to Poor Areas?" *Journal of Public Economics* 93(3–4):512–528.

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).

Devarajan, Shantayanan, Lyn Squire, and Sethaput Suthiwart-Narueput. 1997. "Beyond Rate of Return: Reorienting Project Appraisal." *World Bank Research Observer* 12(1):35–46.

Djebbari, Habiba, and Jeffrey Smith. 2008. "Heterogeneous Program Impacts of PROGRESA." *Journal of Econometrics* 145(1–2):64–80.

Du, Runsheng. 2006. *The Course of China's Rural Reform*. Washington, DC: International Food Policy Research Institute.

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.

Galasso, Emanuela, and Martin Ravallion. 2005. "Decentralized Targeting of an Anti-Poverty Program." *Journal of Public Economics* 89(4):705–27.

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):128–42.

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, and Jeffrey Smith. 1995. "Assessing the Case for Social Experiments." *Journal of Economic Perspectives* 9(2):85–110.

Heckman, James, L. Lochner, and C. Taber. 1998. "General Equilibrium Treatment Effects." *American Economic Review Papers and Proceedings* 88(2):381–6.

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, Serio Urzua, and Edward Vytlacil. 2006. "Understanding Instrumental Variables in Models with Essential Heterogeneity." *Review of Economics and Statistics* 88(3):389–432.

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.

Lanjouw, Peter, and Martin Ravallion. 1999. "Benefit Incidence and the Timing of Program Capture," *World Bank Economic Review* 13(2):257–74.

Lee, Donghoon. 2005. "An Estimable Dynamic General Equilibrium Model of Work: Schooling, and Occupational Choice." *International Economic Review* 46(1):1–34.

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):506–40.

_____. 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*. Lanham: Rowman and Littlefield.

OECD (Organisation for Economic Co-operation and Development). 2007. *A Practical Guide to Ex Ante Poverty Impact Assessment*. Paris: Development Assistance Committee Guidelines and Reference Series, OECD.

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*, 165–90. New York: Oxford University Press.

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.

_____. 2008. "Evaluating Anti-Poverty Programs." In Paul Schultz, and John Strauss, eds., *Handbook of Development Economics*, vol. 4. Amsterdam: North-Holland.

Ravallion, Martin, and Gaurav Datt. 1995. "Is Targeting through a Work Requirement Efficient? Some Evidence for Rural India." In Dominique van de Walle, and Kimberly Nead, eds., *Public Spending and the Poor: Theory and Evidence*. Baltimore: Johns Hopkins University Press.

Ravallion, Martin, and Shaohua Chen. 2005. "Hidden Impact: Household Saving in Response to a Poor-Area Development Project." *Journal of Public Economics* 89(11–12):2183–204.

Ravallion, Martin, and Dominique van de Walle. 2008. *Land in Transition: Reform and Poverty in Rural Vietnam*. Basingstoke: Palgrave Macmillan and World Bank.

Ravallion, Martin, Dominique van de Walle, and Madhur Gaurtam. 1995. "Testing a Social Safety Net." *Journal of Public Economics* 57(2):175–99.

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(1):208–30.

Rodrik, Dani. 2009. "The New Development Economics: We Shall Experiment, but How Shall We Learn?" In Jessica Cohen, and William Easterly, eds., *What Works in Development? Thinking Big and Thinking Small*, Washington: Brookings Institution Press.

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.

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(2):667–85.

Weiss, Carol. 2001. "Theory-Based Evaluation: Theories of Change for Poverty Reduction Programs." In Osvaldo Feinstein, and Robert Piccioto, eds., *Evaluation and Poverty Reduction*. New Brunswick, NJ: Transaction Publications.