Chapter 21

The Wicked Cases of Education and Climate Change - The Promise and Challenge of Theory-Based Impact Evaluations

Emmanuel Jimenez and Jyotsna Puri

Abstract. *To make progress in achieving the 17 Sustainable Development Goals of the 2030 Agenda, policy makers need to know what works to move the needle of the 169 targets, and then to act on that knowledge. While more is known now than when the Millennium Development Goals were set, there are still important gaps regarding the measured attributable impact of interventions on outcomes. This paper focuses on two key goals—education and climate change/environment—to illustrate the gaps in what has been learned and what still needs to be learned. It assesses why these gaps persist, and how future evaluations might address them.*

Emmanuel Jimenez, International Initiative for Impact Evaluation, 3ie, ejimenez@3ieimpact.org; Jyotsna Puri, Green Climate Fund, jpuri@gcfund.org. The views expressed in this chapter are the authors' own and are not the official views of the organizations the authors otherwise represent.

The role of evaluation in helping the world make progress on the 2030 Agenda for Sustainable Development is well argued by other chapters in this volume, as well as by many other thinkers. But it is not enough to monitor indicators: countries also need to know which policies, programs, and other interventions will be effective in moving the 169 indicators of the 17 Sustainable Development Goals (SDGs) to which they have committed.[1] Indeed, in asking for accountable institutions in Goal 16, the SDGs themselves underscore the importance of evaluation.

One key tool in this arsenal is evaluations that address the issue of attribution convincingly. Theory-based impact evaluations do so through methodologies that measure the effectiveness of interventions by posing a counterfactual question: that is, what would have happened without the intervention? In answering this question, they also answer other questions that are important for both donors and implementers to consider: Did the program or policy make a change (and how do we know if it did)? How much was that change? Would the change have occurred anyway, in the absence of the policy? Could it have been done better? Why did the change occur?

Theory-based impact evaluations measure causal change that can be attributed to an intervention, and use a prespecified theory of change to guide their hypotheses and to explain change. Good theory-based impact evaluations usually have the following components: a theory of change; pre-analysis plans; variables that are measured as objectively as possible, using survey data both at the baseline and end line; good pilots and formative work; a good understanding of outcome(s); SMART[2] indicators; good monitoring data and information on implementation fidelity; a good identification strategy; sufficient data size for statistical confidence; and high-quality analyses that mitigate a multitude of possible biases that may creep in over and above the bias of program placement and selection.

In this chapter, we investigate the present state of evidence and argue that theory-based impact evaluations provide a potential partial solution to answering the critical questions that the SDGs are asking. We show that every year many more of these evaluations are being published than ever before. We also discuss the limitations of current methods employed in theory-based impact evaluations, and argue that there are important gaps in the knowledge base in terms of topics and methods that need to be filled if we are to accomplish the goals of the 2030 Agenda in an evidence-based way.

We take a deep dive into two sectors to assess both the opportunities and the limitations for theory-based impact evaluations. Arguably education and environment are sectors that have posed what are termed "wicked" problems for evaluators (see, e.g., Levin et al. 2012). Interventions in these sectors

[1] See the Organisation for Economic Co-operation and Development website (http://www.oecd.org/dac/evaluation/sustainabledevelopmentgoalsandevaluation.htm) for a description of the 19th Development Assistance Committee meeting, which focused on the implications of the SDGs on evaluation.

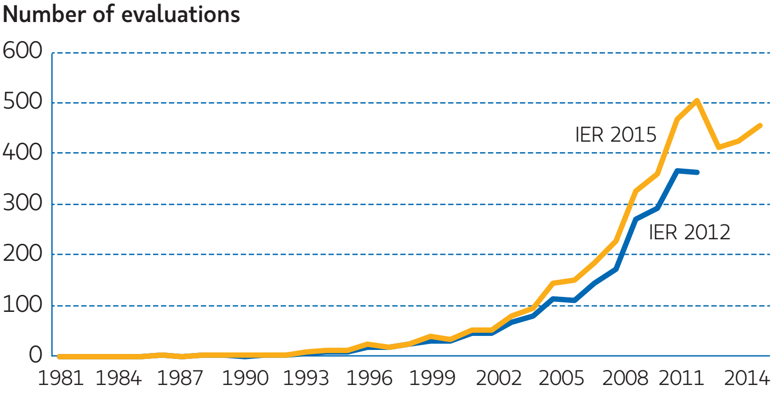
[2] Usually a mnemonic for smart, measurable, achievable, realistic, and time-bound.

resist a single solution because they are applied differently in different contexts. Moreover, the solutions are temporary while they address complex problems that require the use of multiple interventions simultaneously (Schwandt et al. 2016). We discuss how theory-based impact evaluations have tackled these issues and where gaps remain.

IMPACT EVALUATIONS: TRENDS AND EVIDENCE GAPS

The number of theory-based impact evaluations has risen dramatically in the past 20 years. Figure 21.1 shows just one indicator—the number of theory-based impact evaluations of development interventions that are published per year and that take the counterfactual adequately into account. Figures are derived from the 3ie repository,[3] which was initially analyzed by Cameron, Mishra, and Brown (2016) and are currently being updated. In 1995, there were fewer than 50 studies being published per year; by 2015, there were almost 500 and the repository contained more than 4,500 publications. While these figures need to be considered in light of publication lags, they include working papers in the gray literature that have shorter time frames between when the data are collected and when the results become available.

FIGURE 21.1 **Number of development impact evaluations published each year, 1980–2015**



SOURCE: Miranda, Sabet, and Brown 2016.

NOTE: IER = 3ie Impact Evaluation Repository. Data for 2015 are only for the first three quarters.

These numbers alone do not tell us anything about the need or demand for evidence. After all, there are untold hundreds of thousands of public and

[3] The 3ie Impact Evaluation Repository is an index of all published impact evaluations of development interventions.

private programs in more than 150 lower- and middle-income countries in the world. But they can be used as possible indicators of where glaring gaps may exist.

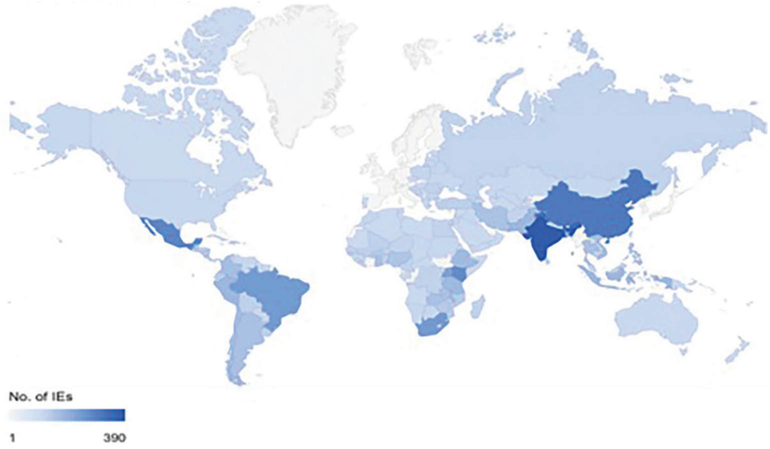
Geographic Gaps

Even with these promising global trends, the density of evidence from rigorous impact evaluations varies widely across countries. Indeed, this is borne out in figure 21.2, which maps countries where studies included in the repository were conducted.

Figure 21.2 shows that countries in Asia (especially the largest countries, China and India), and parts of Latin America (Brazil, Mexico) and East Africa (particularly Kenya and South Africa) have more theory-based impact evaluations than others. This does not mean that these countries need fewer theory-based impact evaluations in the future. In fact, in terms of size of the economy and population, these countries may continue to need many more evaluations (e.g., the number of evaluations per 10 million people in China is about the same as that in Russia; it is more in India but half that of Brazil). But it does show that there are some regions that lag. There are extremely few (none in many countries) evaluations in West Africa, Middle East and North Africa, and Central Asia, the Pacific countries as well as the poorer countries in Latin America and the Caribbean.

The relatively uncovered regions are sites of fragile and conflict states (FCS), where populations are the most vulnerable. Only about 8 percent of published evaluations were done in FCS countries, and almost half of those were in just two countries—Pakistan and Zimbabwe.

FIGURE 21.2 **Impact evaluations by country**



SOURCE: Miranda, Sabet, and Brown 2016.

These geographic gaps pose a significant challenge for the 2030 Agenda, which is a global action plan: all member countries of the United Nations have committed to it. Yet for many there is a remarkably small evidence base that can attribute improved outcomes to interventions.

Thematic Gaps

The majority of all published impact evaluations are in four sectors: health, education, social protection, and agriculture. Again, this is not to say that these sectors are saturated and have no further need for more evidence. But it does point to the key sectors that are consuming vast amounts of public and private expenditure, but are not being evaluated. For example, according to Miranda, Sabet, and Brown (2016), there are very few published theory-based impact evaluations in environment and separately in the energy sectors. To put this in perspective, India's public sector budget allocates a significant portion its budget to energy, and the World Bank has devoted 16 percent of its loans to it as well (World Bank 2016).

Arguably one reason for this lack of impact evaluation-related evidence may be the lack of demand in the sector. Indeed in several sectors, the questions examined by impact evaluations have traditionally not been considered important.[4] How much impact does a road make? Do protected areas reduce deforestation? Do climate change programs work to reduce greenhouse gases (GHGs)? Do children learn once they enroll and attend school? These are all examples of questions that have, until recently, not been considered in time-consuming and resource-intensive evaluations.

Another possible reason for this disparity is that it is believed that it is much more difficult to apply popular techniques of theory-based impact evaluations in some sectors, such as national infrastructure investments, or public finance policy, or practices of good governance, than in other sectors, where the interventions are smaller, easier to isolate, and have identifiable possible counterfactual (or comparison) populations. If so, the question is whether rigorous techniques can be developed to address key issues that obviously have huge implications for human welfare. Such efforts would have to take into account several other reasons why such knowledge gaps persist across sectors. For example, there may be disincentives for political economy reasons, for evaluating already scaled-up investments in sectors like transport, where large amounts of capital, both political and monetary, may have already been sunk (Ravallion 2016).

Aside from the density of evidence across broad sectors, there are gaps in thematic areas that are also of programmatic interest. For example, Puri et al. (2014) found that there were fewer than 50 studies of humanitarian assistance, into which the world has pumped over a trillion dollars. Another report found a single impact evaluation study in the governance

[4] We use the phrase "impact evaluations" and "theory-based impact evaluations" interchangeably. Indeed, we do not believe good impact evaluations can be undertaken without good theories of change.

and transparency of natural resource management in low- and middle-income countries (Puri 2017). Another concern is that in many impact evaluations, the costs of interventions are not analyzed. These trends present huge challenges for informing a comprehensive, global SDG agenda that encompasses almost all sectors in promoting people, planet, and prosperity.

Distributional Gaps

Increasingly, more questions are being asked not only about the overall effects of interventions, but also of their effects on specific groups such as women and girls, vulnerable ethnic groups, the very poor, and so on. But the number of studies that have done such deep distributional analysis is also relatively low.

This gap holds even for sectors where there are more impact evaluations. For example, one of the findings of a recently completed systematic review by 3ie of what works in education is that while studies reported on the average effects on all children, "…[few] studies included in the review provided any analysis of sub-populations, including factors such as sex or socio-economic status" (Snilstveit et al. 2016a, 14). A large part of this is driven by the fact that ensuring that impacts are measured with high statistical confidence for underrepresented groups means that the statistical samples need to be much larger. In another study we estimate that in one case, in order to ensure that results were representative for men *and* women, the sample sizes needed to quadruple (because women are traditionally underrepresented in some economic sectors), which meant a concomitant increase in costs (Puri, Rathinam, and Sarkar 2017).

IMPACT EVALUATIONS: CHALLENGES OF RELEVANCE AND METHODS

As the number of impact evaluations have risen, researchers have learned more about their limitations and how to address them.

The Challenge of Responding to Questions Important for Policy Making

Arguably, theory-based impact evaluations answer several questions that are important for policy making: Does the intervention work? How much? For whom? But sometimes they are just not the right instrument to answer the question. Nowhere is the latter point more salient than when researchers try to fit the question to the method. Simply identifying the underlying theories of change is a complicated enough undertaking, and adding the overall requirement of having the measure attributable to change becomes a daunting task (see, e.g., box 21.1).

The other question is to what degree should research be responding to policy, and whether research is important for its own sake. Theory-based impact evaluations tend to lie at the intersection of research and applied work (see, e.g., Puri, Rathinam, and Sarkar 2017). We argue that impact evaluations

BOX 21.1 **The use of impact evaluations in the evaluation of large, complex climate change programs: How can theories of change help?**

Aided by the Food and Agriculture Organization of the United Nations (FAO), the government of Paraguay is implementing a program to alleviate poverty and help reforest a large part of eastern Paraguay and increase the resilience of approximately 62,000 people. The proposal is to implement cross-cutting programming that meets both mitigation objectives (725,000 tons of carbon dioxide mitigated annually) and adaptation objectives (an expected direct increase in resilience and a reduction in poverty for 62,000 people). This three-phase program is spread out over 10 years and supports components such as environmental conditional cash transfers, cook stoves, and agroforestry programs for households. It aims to simultaneously improve the legislative and institutional frameworks mainly of forestry, environmental, and energy regulating entities. The overall objective of the program is to improve the resilience of poor and extremely poor households vulnerable to the impacts of climate change in environmentally sensitive areas of eastern Paraguay.

The implicit theory of change of the program is that (1) once authorities have the requisite funds and approvals, they will be able to set up environmental conditional cash transfer (E-CCT) payment systems that piggyback on existing cash transfer systems that already target poor and vulnerable communities through automated banking systems; (2) households will be targeted successfully; (3) as a consequence of the incentive of E-CCTs, households will start to build and invest in agroforestry systems (for which they will be paid for inputs and provided with technical assistance) they would not otherwise have; (4) households' agroforestry systems will be measured and detectable, which will then trigger payments to them; and (5) forest cover and degradation in eastern Paraguay will be reduced as a consequence and climate change mitigation will occur. There are several assumptions here, including assuming that households will be able to take the surplus produce from the agroforest systems to local and regional biomarkets, and that they will be able to earn incomes from these which will also reduce their income poverty and therefore increase their resilience.

Clearly, all of these statements require either strong previously produced evidence or smaller evaluative tests to understand whether the linkages are working, and whether the overall effects of the program will be achieved.

bridge a very important gap, in that they apply science and rigor to questions that have previously been hand-waved.

The Challenge of Complexity

Complexity poses a substantial challenge to impact evaluations. Many programs involve a multitude of sectors: for example livelihood programs include interventions in water provision, sanitation, income-generation activities, and health. This usually means that causal pathways are not direct, are cross-linked, and are nonlinear. Separately, it also means that there are a multitude of sectors that every program is aiming to target. Arguably this reduces the incentive for any one sector team to invest in impact evaluations. Moreover, because there are intersectoral links and feedback loops, it becomes harder for impact evaluations to answer the "why" questions once the "how much" questions have been answered. Woolcock (2013) frames this challenge by citing three specific challenges that randomized control trials (RCTs) are unable to deal with. He cites the challenges of "causal density," "implementation capability," and "reasoned expectations" as being key features of complex systems that also make it difficult for RCTs to be used for understanding the overall impact of development interventions.

Another aspect of complexity is the measurement of relevant outcomes. For example, test scores may be an important indicator of performance in an education project, and RCTs may indeed be able to measure these well. It may also be possible to create good indicators and to include these in pre-analysis plans. But student stress may be an unintended consequence of these score-enhancing programs. There are two difficulties here: the inability to prespecify all possible consequences in a protocol, and the difficulty in measuring student stress caused by these programs.

Another example of the measurement challenge is climate projects that are aimed at increasing adaptation. A recent survey by the Overseas Development Institute found that there are at least 43 different frameworks for defining and understanding climate adaptation (ODI 2016). Again, because climate change programs also typically incorporate poverty alleviation and equity as a primary objective, these causal chains become very difficult to identify. As Levin et al. (2012) point out, it also becomes more likely that a specific solution to one development challenge creates a new problem for another one. Road building is touted as one such development solution that has clear (negative) implications for forest cover and biodiversity.5 Additionally, confounding features of programs make it difficult to identify and measure the key change the program is seeking to bring about. Given these challenges, designing impact evaluations becomes even more difficult.

[5] Cropper, Puri, and Griffiths (2001) and Puri (2016) discuss the exceptions to this rule.

The Challenge of External Validity

Limited external validity is another limitation of theory-based impact evaluations. Other authors have raised this concern as an important detraction from impact evaluations (Basu 2013; Pritchett and Sandefur 2014; Woolcock 2009). These are important concerns, and impact evaluations will need to respond to them by using new tools. It is of course true that non-impact evaluations are typically not externally valid either. We argue that in this case theory-based impact evaluations, because they are able to articulate the theories behind overall interventions and also provide statistical estimates with confidence intervals, become easier to aggregate through meta-analysis. Although limited, in these cases it is easier to say something about "net" or "aggregate" effect sizes (see Snilstveit et al. 2016b; Waddington and White 2014).

The knowledge gaps and methodological challenges discussed above pose challenges for evaluating the effects of interventions that will help countries address the 2030 Agenda. But they are not insurmountable. In this and the next section, we discuss these challenges and how evaluators have tried to address them in the "wicked" sectors of education and climate change.

WICKED SECTOR: EDUCATION

According to UNESCO's post-2015 Global Education Monitoring report (UNESCO 2015), in order to achieve the ambitious SDG targets for education by 2030, the spending per primary school student in low-income countries needs to be double the current level of spending. The International Education Commission calls for total spending in education to *triple* from its present $1 trillion. But more funding is not sufficient for addressing the learning crisis: resources need to be directed to programs that work. There are a large number of reports about education, but there are relatively few that address attribution directly. 3ie recently completed a comprehensive systematic review of the effectiveness of 21 different types of education programs on children's school enrollment, attendance, drop-out rates, completion, and learning outcomes (Snilstveit et al. 2016a). It included evidence covering more than 16 million children across 52 countries, participating in 216 education programs in 52 low- and middle-income countries. The findings from this study can help inform decisions about effective strategies for achieving the education targets.

The review drew on evidence from 238 impact evaluations and 121 qualitative research studies and process evaluations. Interventions such as cash transfers, structured pedagogy, and computer-assisted learning programs were studied extensively. For other programs, such as school-based health, information to children, teacher interventions, remedial education, and school-day extension, the evidence is more limited. Significant investments are being made for funding interventions in understudied areas such as teacher-related programs. There is an urgent need for generating more evidence to help in informing funding decisions.

The education sector mirrors global variation in the availability of evidence. The greatest number of studies was identified in Latin America and the Caribbean (87); Sub-Saharan Africa (59); and South Asia (51). Countries where several studies have been conducted include Brazil, Chile, China, India, Kenya, Mexico, South Africa, and Uganda. Evidence is limited or nonexistent for many countries in Sub-Saharan Africa, and for several countries with large populations, such as Bangladesh, Indonesia, and Nigeria.

Aside from inadequate thematic and geographic coverage of some important interventions, the usefulness of impact evaluations for the 2030 Agenda faces another challenge. Many education interventions are methodologically "wicked" to evaluate. Three aspects are particularly important to consider: the logical chain of intervention to results; context; and implementation.

The Logical Chain from Interventions to Results

While some interventions have a relatively simple logical chain from intervention to results, such as the provision of textbooks on learning, or of scholarships on school participation, many others are characterized by causal density. This means that the interventions are "…highly transaction intensive, require considerable discretion by implementing agents, yield powerful pressures for those agents to do something other than implement a solution, and have no known (ex ante) solution" (Woolcock 2013).

It is thus not surprising that interventions that have a direct and simple link to the desired outcome—short results chains—are more effective. For example, cash transfer programs were the most effective intervention to boost school attendance. Where the outcomes of any one intervention are conditioned by the effectiveness of other interventions that may be beyond the scope of the program, the results tend to be more mixed. For example, in contrast to their effect on school participation, cash transfers have very little effect on learning outcomes as measured by mathematics or reading scores. This may not be surprising, given that most of the programs were conditioned on school participation and attendance, not on test performance. But the fact that learning outcomes were not significantly affected may also be a reflection of the low quality of schools that children were incentivized attend. In Colombia, for example, school vouchers (which are effectively conditional cash transfers) had no effect on learning outcomes if they were limited to government schools, but had a positive effect in those areas where the recipients were able to use them for entry into private schools, most of which were perceived to be of higher quality (Barrera-Osorio et al. 2011).

Another example is the need to address the incentives of the most important actor in affecting students in classrooms—the teacher—in almost any intervention. Some of the teachers who were delivering the Reading to Learn program in Kenya chose not to accept the class materials because they considered them difficult to master. This may have been one of the reasons that the program did not improve children's performance in written and oral literacy exams. Similarly, the evidence on computer-assisted learning programs suggests that while the implementation of training for teachers is an issue, program designs need to also consider teacher workloads, as well as their attitudes and motivation for making radical changes in the way they teach. School-based health programs also require teachers to participate in program delivery. Hence, programs need to consider whether this is increasing the workload for teachers and disrupting the regular class routine.

Baseline Conditions and Local Context

Many successful interventions have tailored their design well to the existing human and social capital of specific contexts. This is particularly important for interventions aimed at children and households, and those aimed at improving governance.

School feeding programs, for example, have had the largest effect in areas characterized by high levels of food insecurity, malnutrition, and low school attendance. The effects have been much smaller in better-off areas where enrollment was already high, and malnutrition less of an issue. The school feeding program in Guyana, for instance, was implemented at a time when there was a documented increase in food insecurity for poor families. Not surprisingly, the program had large positive effects on school participation and learning. However, in Chile, the effects of the program on school participation were found to be small or nonexistent. In this case, the program was implemented at a time when extreme malnutrition had been eliminated, and enrollment rates were already high.

The baseline level of social capital has been found to be more important in interventions aimed at improving the system of governance of schools. School-based management and community-based monitoring had the best take-up in settings with high levels of social capital and a tradition of local participation. In the Philippines, where the effects of school-based management were consistently positive, qualitative evidence suggests that parents and communities were willing and able to make basic decisions about schooling when given the opportunity to do so. In contrast, results in most other contexts were disappointing. Evidence from Niger and Gambia pointed to low social and human capital as an important constraint for school-based management programs. Programs that rely on parental engagement for successful implementation may be better targeted in contexts where there is sufficient social and human capital to be able to hold other stakeholders accountable. For instance, where school committees are educated, or have experience in another community organization, parental monitoring of teacher attendance is likely to increase in response to the grant. Where these conditions are not met, programs may have a higher chance of success if there is a strong capacity-building component that is focused on facilitating community involvement. More generally, when parental engagement is a key part of the theory of change of a program it is important to assess the local capacity to engage in the way assumed by the theory of change. Programs could then be designed to account for any deficit in social and human capital.

It is therefore imperative that decision makers obtain accurate baseline information at the design stage of the program. This is required in order to tailor new programs to target the main constraints and achieve better outcomes.

Capacity to Implement

The success, or more often the failure, of a program has often been attributed to the way the program has been implemented. Issues related to implementation have frequently been reported for a range of programs. The effect sizes of some programs were much smaller due to implementation problems. For example, in Kenya's and Uganda's Reading to Learn, as well as in Mali's Read, Learn, Lead programs, school materials and other tools were not delivered in a timely manner, which may partly explain why they had no effects, or very small ones, compared to the overall average for structured pedagogy on learning outcomes (Snilstveit et al. 2015). Similarly, the distribution of textbooks to students was found to be lower than intended in the case of a few programs that were providing school materials.

Several computer-assisted learning programs have faced issues such as insufficient, damaged, and dysfunctional equipment, lack of Internet access, and software not being compatible with hardware. Insufficient training of teachers is another issue that has been brought up as a challenge for several programs, including computer-assisted learning. Implementation issues, particularly with respect to the transfer of funds affected the success of several school-based management programs. Grants were not disbursed as intended, and significant delays were reported for several programs. Finally, unforeseen circumstances such as epidemics and conflicts have also delayed the implementation of education programs.

In most cases, these issues have cropped up due to the lack of capacity for implementation at various levels of the supply chain. In some cases, the inability to ensure a sustained and timely supply of resources has affected the effectiveness of programs. The difficulty in implementation is also often seen in programs that include a range of activities, and that have ambitious goals and long causal chains. This leaves a lot of room for implementation failure. In contexts where there is limited capacity to implement it may be necessary to give up on some of the objectives in the interest of making the program capable of implementation.

Summary Implications

All of these complications point to the need for better-designed impact evaluations: those that study multiple options (or arms) to test different combinations of interventions would be greatly beneficial in addressing the causal complexity of some education interventions. But the examples above also point to the need for mixed methods in evaluation. Rigorous case studies, such as in Woolcock (2013), as well as incremental approaches to learning as in Andrews, Pritchett, and Woolcock (2012) would be one way to approach this. Finally, rigorous estimates of effects must be accompanied by equally rigorous studies of implementation.

WICKED SECTOR: CLIMATE CHANGE

In this section we discuss the overall strengths and limitations of using theory-based impact evaluations in climate change programs and policies. It is clear that the international policy arena has parsed climate change into several components, perhaps recognizing their overwhelmingly large reach and scope. So, for example, there are different conferences of parties (COPs) for climate and for forestry. Organizations and funding are also largely segregated into three different areas or sectors—mitigation, adaptation, and forestry. The United Nations Framework Convention on Climate Change recognizes these areas. Therefore, we define climate change activities and sectors as all those that help to reduce or stabilize GHG emissions, and that help to increase adaptation to climate change and its resulting uncertainties and weather extremes.

In the mitigation category, a host of types of policies and programs are included—these include policies and programs that increase access to and the use of low-emission energy and power generation; programs that increase access to and the use of low-emission transport; energy-efficient buildings, cities, and industries; and programs and policies that aim to increase sustainable land use and forest management, including reducing emissions from deforestation and forest degradation, or REDD+, programs. In the adaptation category, the range of policies and programs includes those that increase resilience and enhance livelihoods of vulnerable people, communities, and regions; that increase resilience of health and well-being; that increase food and water security; that increase the resilience of infrastructure and built environment to climate change threats; and that increase the resilience of ecosystems.

Challenges of Evaluating Climate Change Action

Evaluations of programs and policies that deal with climate change encounter some of the challenges laid out in the section on education, and others as well. First there is the challenge of distal impacts. Climate change mitigation takes time (and scale): assessing overall contribution to climate change mitigation requires long time horizons. With theory-based impact evaluations, some of this is dealt with by underlying theories, mapping outcomes, and assessing efficacy and program success (see, e.g., box 21.1). The overall question related to understanding and measuring change, however, remains a challenge. Indeed, an evidence gap map examining the effects of land use policies on mitigation (Snilstveit et al. 2016b) found that, although there were 221 studies that rigorously looked at the impacts of land use policies and interventions on outcomes such as tree cover, livelihoods, and health, there were *no* evaluative studies that linked these, in an attributable way, in developing countries, to GHG emissions. This not only speaks to the difficulty of waiting for long periods of time for these impacts to show: it also underscores the difficulty of measuring GHG emissions. The other difficulty in these programs is that in order for there to be a *measurable* effect on even GHG emissions, programs have to account for “leakages,” that is, the likelihood that mitigation programs in one area may lead to the displacement or movement of emission activities to other areas. Impact evaluations therefore have to cover large areas, in order to ensure that there is a net effect on GHGs. This public good nature of climate change action imposes large transaction costs, but it also means that impact evaluations that aim to measure attributable change have to focus on large-scale action, and this may not always be possible (box 21.2).

BOX 21.2 A large-scale mitigation program: An example of solar home systems

Bangladesh's solar home systems (SHS) program—supported by the World Bank, GIZ, KfW, the European Union, the Inter-American Development Bank, and the multidonor Global Partnership for Output-based Aid trust fund—aims to provide energy for poor and vulnerable households. To gauge its uptake and effects, an evaluation was undertaken 10 years after its inception. The findings revealed a complex set of factors at play. To begin with, by 2013, only 10–12 percent of off-grid households had access to SHS off-grid devices, and diffusion rates were low: on average, a maximum of one-third of eligible households had adopted SHSs. The households that adopted the devices were, on average, much richer (80 percent higher incomes than non-adopting households) and better educated, with high percentages of non-agricultural income and a higher level of household assets. More than 78 percent of the adoption had only occurred during the last three years of the program. Despite their SHS adoption, most households continued to depend on traditional sources of energy. While there was some evidence of a substitution effect with SHS replacing kerosene, SHS households overall consumed more energy compared to non-SHS households—indicating that the income effect was stronger than the substitution effect. An important factor influencing adoption is the cost (including interest cost) and maintenance of SHS devices. But over a quarter of those taking out loans for the SHS devices—which are sold on credit, with loans provided for three years with a flat interest rate of 6 percent—were late in their repayments. Clearly for the program to conclude that it has been effective in achieving its long-term goal, given that the magnitude of change in overall emissions will be important, evaluation will need to measure the income effect and substitution effect over time.

SOURCE: Adapted from Asaduzzaman et al. 2013.

The second challenge for climate change evaluations is that most climate change projects have multiple objectives. This means that other than feedback loops and backward and forward links, most climate change projects are not just planned and implemented to maximize impact on climate change, but to simultaneously affect social, economic, health, and agricultural objectives. This means that the strength of links in a theory of change are frequently not the same, that they intersect and impacts mostly depend on the efficacy of several links being realized. This makes impact evaluations—which assume that a single intervention will lead to the overall impact, all other things being held constant—difficult to plan, implement, and realize in this space.

Third, climate and the environment are inherently public goods. This means that the overall impact of interventions is not individually determined by the successful implementation of one project over one discrete area.

Rather, in a twist on the problem of tragedy of commons, it is characterized by the problem of large numbers with small payoffs. This scale problem has two implications for impact evaluations. First, it means that impacts do not show unless there are a large number of agents. Second, they do not show unless a large number of agents are successfully undertaking these actions. Therefore, impact evaluations of climate change programs and policies in most cases have to concentrate on measuring attributable change at the outcome level. Furthermore, in most cases, although small programs may themselves be successful, we still may not see any changes in overall climate change-related impacts: this is true for both mitigation and adaptation programs. In some cases this means that small climate- related impacts have to accumulate until we are able to confidently detect and witness a change. As Bamberger, Rao, and Woolcock (2016) and Woolcock (2009) explain it, the impact function for climate change programs and policies may be nonlinear, or they may be horizontal straight lines before we see any impact. Examining and understanding the role that scale plays in identifying and measuring impact with statistical confidence while planning impact evaluations is therefore very critical.

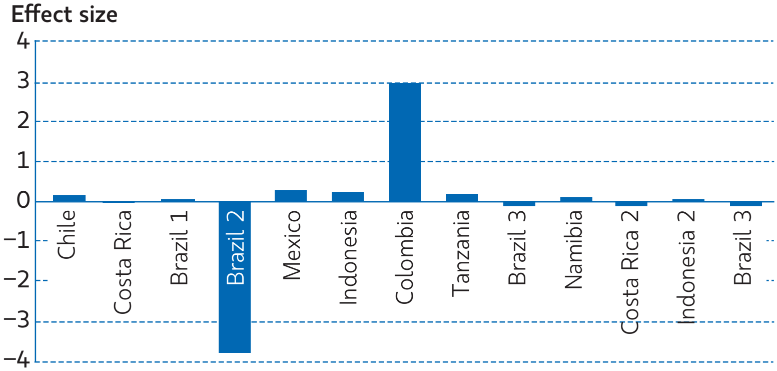
Credible and high-quality impact evaluations are critically dependent on defining the appropriate questions and "system boundaries," that is, defining the type and nature of the interventions that will be examined through impact evaluations. This in turn implies that impact evaluations are only able to examine a shortlist of interventions that have been defined using a variety of designs, such as factorial designs and pipeline designs. They are, by themselves, unable to compare interventions that have not been shortlisted. Arguably though, theory-based impact evaluations can deal with these through doing good initial formative work and specifying unintended consequences, and by undertaking rigorous qualitative work along with data collection that can help inform areas that previous theories may have been blind to (see, e.g., Rao and Woolcock 2002). We believe this is important to understand and concede, primarily because theory-based impact evaluations also have the advantage of lending themselves to systematic reviews with statistical meta-analyses that help us understand aggregate average effects, but also help us view the distribution of these effects; identify and analyze outliers; and examine other effects, such as "dose-response" pathways. In the next section, we will review some learning from impact evaluations in this area, and discuss new areas that theory-based impact evaluations should focus on, given the challenges and gaps.

What Are We Learning?: Recent Evidence from Theory-Based Impact Evaluations and Systematic Reviews

Theory-based impact evaluations have helped us understand the amount of change that environmental programs are bringing about. In figures 21.3 and 21.4, we show an illustrative summary of the magnitude of impact that several impact evaluations are able to measure in forestry programs.[6]

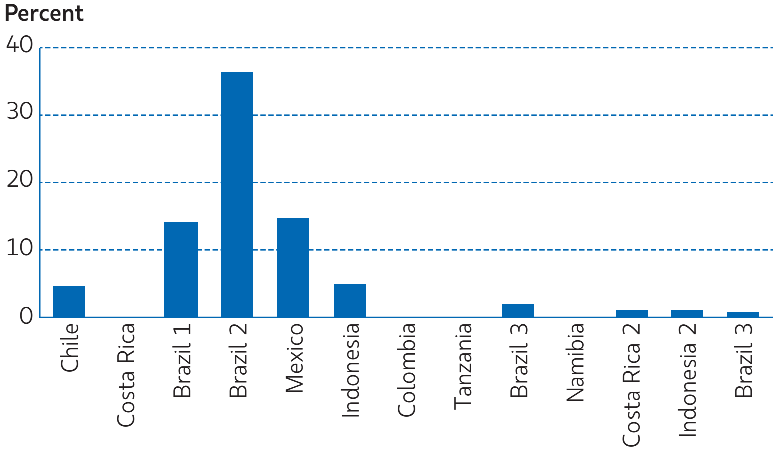
[6] See PLOS Collections (2016) for a summary of statistics from different forestry programs.

FIGURE 21.3 **Standardized effect sizes from forestry programs**



SOURCE: Adapted from Börner et al. 2016.

FIGURE 21.4 **Estimate reduction in forest loss as a consequence of forestry programs**



SOURCE: Adapted from Börner et al. 2016.

Bias. The important thing to note is that impact evaluations help deal with the problem of endogeneity and placement bias. Cropper, Puri, and Griffiths (2001) and Chomitz and Gray (1996) account for the attributes of plots where protected areas tend to be sited or located. Since these are areas that have low agricultural productivity and are likely to be remote, it is clear that any naïve estimate that does not consider this selection bias is likely to have extremely biased results.

Targeting. Impact evaluations can also help us understand the effectiveness of targeting: Are programs really reaching the populations that they need to, and are these the populations that programs most need to target? Alix-Garcia, Shapiro, and Sims (2012) found that the countrywide Payment for Ecosystem Services program (PSA) in Mexico, with a budget of more than $5 million, was quite successful in targeting households that were eligible for the program. In contrast, Azofeifa et al. (2007) found that in Costa Rica, the PSA program did not target those locations that were most likely to change land use. As a consequence there were very small changes in forest cover caused by the PSA program.

Subgroup effects. Impact evaluations can also help to address questions of equity and heterogeneous impacts. Somanathan, Prabhakar, and Mehta (2009) have shown that after accounting for potential selection and placement bias, community-managed forests performed better in raising crown cover by 12–16 percent when compared to unmanaged commons, but only for forests of broadleaf trees, not pine trees. Understanding the effects on subgroups, however, requires that sample sizes be selected in such a way that they are representative for the subgroups of interest.

Comparing different kinds of programs. Many studies have examined programs that engage communities and compare their effectiveness with the status quo, such as government managed systems, or unmanaged systems. For example, Tachibana and Adhikari (2009) showed that in Nepal, community co-managed forests recover much more quickly than forests where communities are solely managing their forests. And Cropper, Puri and Griffiths (2001) found that protected areas are less effective in protecting forests than wildlife sanctuaries by themselves, perhaps because the latter have more resources devoted to them.

Are we doing the right things? In our view, one of the key questions that impact evaluations should answer is, “Are the right things being done?” A study by Wynes and Nicholas (2017) found that of the top four mitigation actions that individuals can undertake to reduce GHGs, only two are discussed in high school textbooks. The other two actions are completely ignored. If we are to focus policy and action on the most effective actions, evaluations need to start asking the question of whether the right strategies are being pursued programmatically, rather than evaluating only the implementation of policies. While traditional evaluations have mostly been unsuccessful in this area, impact evaluations can help us respond to this overall question.

Trade-offs. A relatively but clearly an important question in climate change is examining any potential trade-offs between economic outcomes on one side and environmental outcomes on the other. This is particularly important in the case of programs that aim to reduce the consequences of development on the environment. A good example is provided by the Alix-Garcia et al. (2013) study of Oportunidades, a conditional cash transfer program, and its consequences for forests. The authors of this study found that forests were detrimentally affected as a consequence of a cash transfer program, and that the theory-based impact evaluation helped to measure the magnitude of this effect. This is important, because program managers and policy makers can then measure the magnitude of this effect and make policy decisions accordingly.

CONCLUSIONS: ARE THERE COMMON OPPORTUNITIES TO ADDRESS THE GAPS?

Theory-based impact evaluations have been used across the development and humanitarian sectors to inform the effectiveness of programs. This includes investigating the best ways to deliver humanitarian assistance (see, e.g., Doocy and Tappis 2016; Puri et al. 2017); examining the effectiveness of self-help groups in empowering women through microfinance (see, e.g., Brody et al. 2016), community-driven development (King, Samii, and Snilstveit 2010), sanitation programs (Buck et al. 2017), farmer field schools (Waddington and White 2014), agricultural insurance (Barooah, Kaushish, and Puri 2017), reducing poverty (Banerjee et al. 2015), and day care programs (Leroy, Gadsden, and Guijarro 2012). In so doing, they have helped to reduce ambiguity in our knowledge of the effectiveness of development programs.

However, many challenges remain. First, theory-based impact evaluations have not succeeded in meeting the methodological challenges discussed in this chapter. Additionally, it is clear that theory-based impact evaluations have not really leveraged the data revolution. The methods traditionally employed in theory-based impact evaluations have largely remained the same, predicated on the assumption that data are scarce and infrequent. Advances are being made with machine learning that use frequent, high density, spatially disaggregated data to understand especially the heterogeneity of impacts, but they are making their way only very slowly into theory-based impact evaluations.

Second, theory-based impact evaluations have largely shied away from meso-level or macro investigations. Causal identification through the use of controls or comparison groups remains a challenge here. Some studies are using innovative methods such as synthetic controls and machine learning (see, e.g., Acemoglu et al. 2013; Sills et al. 2015). Others are still venturing into meso-level investigations using regional controls (see, e.g., Bos et al. 2017). These applications, however, remain infrequent.

Third, theory-based impact evaluations have stayed mostly quiet on systems thinking and on understanding what changes institutions. Methodologies have been limited in this space. This is important because most agencies, especially environmental agencies such as the Global Environment Facility, the Climate Investment Fund, and the Green Climate Fund are aiming to achieve "transformational change." An important characteristic of transformational change is being able to detect and measure systems change. This is an important policy imperative not just for environment-related organizations, but for the development sector as a whole, and it will be useful for the evaluation community to engage closely with applied academics to explore methodological options that help to identify and measure causal contribution in this area.

As recent papers have demonstrated (PLOS Collections 2016), the size of causal change differs dramatically depending on the spatial resolution of data: the less the resolution, the greater the imprecision, but the higher the resolution, the greater the heterogeneity in impacts across intervention sites. Arguably, therefore, it may be even more important to measure the cost-effectiveness of programs and projects (PLOS Collections 2016). Unfortunately this is not something that a lot of evaluations do. The absence of data on costs of implementation is usually cited as a reason for this absence of analysis. But we believe it as important to understand overall effect size as it is to measure cost effectiveness. Unfortunately, as we have shown in one of our papers, there are very few studies that examine cost effectiveness (Puri et al. 2016).

We also believe that evaluations that contribute to implementation science by examining *how* programs may be implemented are far more important than measuring the overall effects of programs. Although other sectors, such as nutrition and health, have long held this as an important area of exploration (see, e.g., Menon et al. 2014), impact evaluation techniques, especially those related to causal identification, have not found widespread use. Comparing different delivery mechanisms and how effective they are in realizing results is especially important. An example of this can be seen with Doocy and Tappis (2016), where the authors compared the effectiveness of cash transfers versus food transfers, versus in-kind transfers in humanitarian contexts. Within this class of research we also recommend using impact evaluations to examine "last mile" questions. Most programs assume that good implementation leads to good results. However, as has been most recently explored by the behavioral insights literature, good implementation is a necessary, but not a sufficient condition for success in development programs. These last mile problems have been examined in the context of the adoption of new technologies (e.g., Burwen and Levine 2012) or for new instruments (Barooah, Kaushish, and Puri 2017). Most programs fail because they presume that good monetary incentives are in themselves sufficient to ensure results. However, the literature on behavioral insights has now shown us that these assumptions are unrealistic.

Despite all of these challenges, we remain sanguine. Theory-based impact evaluations have been able to answer many difficult questions. They have helped policy makers and evaluators understand and measure overall results, and deal with a variety of biases while understanding the impact of development assistance. They have arguably helped to turn the tide in international assistance by providing comparisons of the effectiveness of different programs, that for long periods of time had been accepted as being successful and useful. Theory-based impact evaluations have provided us with a method for comparing strategies as well understanding their relative impact, while developing a systematic way to aggregate effects and understand average impact. Clearly the field is still in its infancy, though, and new, customized methodological advances will be required if we are to answer the questions that are relevant to the policy community.

ACKNOWLEDGMENTS

The authors would like to acknowledge, without implicating them, the comments received from Rob D. van den Berg and the research assistance provided by Swarnakshi Luhach.

REFERENCES

Acemoglu, D., S. Johnson, A. Kermani, J. Kwak, and T. Mitton. 2013. "The Value of Connections in Turbulent Times: Evidence from the United States." *Journal of Financial Economics* 121 (2): 368–91.

Alix-Garcia, Jennifer M., Elizabeth N. Shapiro, and Katharine R.E. Sims. 2012. "Forest Conservation and Slippage: Evidence from Mexico's National Payments for Ecosystem Services Program." *Land Economics* 88 (4): 613–38. doi:10.3368/le.88.4.613.

Alix-Garcia, Jennifer, Craig McIntosh, Katharine R.E. Sims, and Jarrod R. Welch. 2013. "The Ecological Footprint of Poverty Alleviation: Evidence from Mexico's Oportunidades Program." *Review of Economics and Statistics* 95 (2): 417–35. doi:10.1162/REST_a_00349.

Andrews, Matthew R., Lant Pritchett, and Michael Woolcock. 2012. "Escaping Capability Traps through Problem-Driven Iterative Adaptation (PDIA)." HKS Faculty Research Working Paper Series RWP12-036, John F. Kennedy School of Government, Harvard University.

Asaduzzaman, M., Mohammad Yunus, A.K. Enamul Haque, A.K.M. Abdul Malek Azad, Sharmind Neelormi, and Md. Amir Hossain. 2013. "Power from the Sun: An Evaluation of Institutional Effectiveness and Impact of Solar Home Systems in Bangladesh." Bangladesh Institute of Development Studies, Dhaka.

Azofeifa, Arturo Sanchez, Alexander Pfaff, Juan Andres Robalino, and Judson P. Boomhower. 2007. "Costa Rica's Payment for Environmental Services Program: Intention, Implementation, and Impact." *Conservation Biology* 21 (5): 1165–73. doi:10.1111/j.1523-1739.2007.00751.x.

Bamberger, Michael, Vijayendra Rao, and Michael Woolcock. 2016. "Using Mixed Methods in Monitoring and Evaluation: Experiences from International Development." Policy Research Working Paper No. 5245, World Bank, Washington, DC.

Banerjee, Abhijit, Esther Duflo, Nathanael Goldberg, Dean Karlan, Robert Osei, William Parienté, Jeremy Shapiro, Bram Thuysbaert, and Christopher Udry. 2015. "A Multifaceted Program Causes Lasting Progress for the Very Poor: Evidence from Six Countries." *Science* 348 (6236). doi:10.1126/science.1260799.

Barooah, Bidisha, Bharat Kaushish, and Jyotsna Puri. 2017. "Understanding Financial Risks for Smallholder Farmers in Low- and Middle-Income Countries: What Do We Know and Not Know?" 3ie Scoping paper 9, International Initiative for Impact Evaluation, New Delhi.

Barrera-Osorio, Felipe, Marianne Bertrand, Leigh L. Linden, and Francisco Perez-Calle. 2011. "Improving the Design of Conditional Transfer Programs: Evidence from a Randomized Education Experiment in Colombia." *American Economic Journal: Applied Economics* 3 (2): 167–95. doi:10.1257/app.3.2.167.

Basu, Kaushik. 2013. "The Method of Randomization and the Role of Reasoned Intuition." Policy Research Working Paper No. 6722, World Bank, Washington, DC.

Börner, Jan, Kathy Baylis, Esteve Corbera, Driss Ezzine-de-Blas, Paul J. Ferraro, Jordi Rosés, Renaud Lapeyre, U. Martin Persson, and Sven Wunder. 2016. "Emerging Evidence on the Effectiveness of Tropical Forest Conservation." *PLOS ONE*. doi:10.1371/journal.pone.0159152.

Bos, Astrid B., Amy E. Duchelle, Arild Angelsen, Valerio Avitabile, V. de Sy, Martin Herold, Shijo Joseph, et al. 2017. "Comparing Methods for Assessing the Effectiveness of Subnational REDD+ Initiatives." *IOP Science: Environmental Research Letters* 12 (7). doi:10.1088/1748-9326/aa7032.

Brody, Carinne, Thomas de Hoop, Martina Vojtkova, Ruby Warnock, Megan Dunbar, Padmini Murthy, and Shari Dworkin. 2016. "Economic Self-Help Group Programs for Improving Women's Empowerment." 3ie Systematic Review 23, International Initiative for Impact Evaluation, London.

Buck, Emmy De, Hans Van Remoortel, Karin Hannes, Thashlin Govender, Selvan Naidoo, Bert Avau, Axel Vande Veegaete, et al. 2017. "Promoting Handwashing and Sanitation Behaviour Change in Low- and Middle-Income Countries: A Mixed-Method Systematic Review." Systematic Review 36, International Initiative for Impact Evaluation, London.

Burwen, Jason, and David Levine. 2012. "A Rapid Assessment Randomized-Controlled Trial of Improved Cookstoves." *Journal of the International Energy Initiative* 16 (3).

Cameron, Drew, Anjini Mishra, and Annette Brown. 2016. "The Growth of Impact Evaluation for International Development: How Much Have We Learned?" *Journal of Development Effectiveness* 8 (1): 1–21. doi:10.1080/19439342.2015.1034156.

Chomitz, Kenneth M., and David A. Gray. 1996. "Roads, Land Use, and Deforestation: A Spatial Model Applied to Belize." *World Bank Economic Review* 10 (3): 487–512. doi:10.1093/wber/10.3.487.

Cropper, Maureen, Jyotsna Puri, and Charles Griffiths. 2001. "Predicting the Location of Deforestation: The Role of Roads and Protected Areas in North Thailand." *Land Economics* 77 (2): 172–86. doi:10.2307/3147088.

Doocy, Shannon, and Hannah Tappis. 2016. "Cash-Based Approaches in Humanitarian Emergencies." 3ie Systematic Review Report 28, International Initiative for Impact Evaluation, London.

King, Elisabeth, Cyrus Samii, and Birte Snilstveit. 2010. "Interventions to Promote Social Cohesion in Sub-Saharan Africa." Synthetic Review 002, International Initiative for Impact Evaluation.

Leroy, J.L., P. Gadsden, and M. Guijarro. 2012. "The Impact of Daycare Programs on Child Health, Nutrition and Development in Developing Countries." 3ie Systematic Review 7, International Initiative for Impact Evaluation, London.

Levin, Kelly, Benjamin Cashore, Steven Bernstein, and Graeme Auld. 2012. "Overcoming the tragedy of super wicked problems: constraining our future selves to ameliorate global climate change." *Policy Sciences* (Springer US) 45 (2): 123-152. doi:10.1007/s11077-012-9151-0.

Menon, Purnima, Namukolo M. Covic, Paige B. Harrigan, Susan E. Horton, Nabeeha M. Kazi, Sascha Lamstein, Lynnette Neufeld, Erica Oakley, and David Pelletier. 2014. "Strengthening Implementation and Utilization of Nutrition Interventions through Research: A Framework and Research Agenda." *Annals of the New York Academy of Sciences.* doi:10.1111/nyas.12447.

Miranda, Jorge, Shayda Sabet, and Annette Brown. 2016. "Is Impact Evaluation Still on the Rise?" *International Initiative for Impact Evaluation* blog.

ODI (Overseas Development Institute). 2016. "Analysis of Resilience Measurement Framework and Approaches." Brief.

PLOS Collections. 2016. "Measuring Forest Conservation Effectiveness." PLOS.

Pritchett, Lant, and Justin Sandefur. 2014. "Context Matters for Size: Why External Validity Claims and Development Practice Do Not Mix." *Journal of Globalization and Development* 4 (2): 161–97. doi:10.1515/jgd-2014-0004.

Puri, Jyotsna. 2017. “Development Priorities: What Are We Learning about What We Know?” Presentation at the African Evaluation Association Meetings, Kampala, March 27–31.

Puri, Jyotsna, Anastasia Aladysheva, Vegard Iversen, Yashodhan Ghorpade, and Tilman Brück. 2014. “What Methods May Be Used in Impact Evaluations of Humanitarian Assistance?” 3ie Working Paper 22, International Initiative for Impact Evaluation, New Delhi.

———. 2017. “Can Rigorous Impact Evaluations Improve Humanitarian Assistance?” *Journal of Development Effectiveness* 9 (4): 519–42. doi:10.1080/19439342.2017.1388267.

Puri, Jyotsna, Francis Rathinam, and Ritwik Sarkar. 2017. “Real World Impact Evaluations—What Are We Learning.” Draft Paper, International Initiative for Impact Evaluation, New Delhi.

Puri, Jyotsna, Megha Nath, Raag Bhatia, and Louise Glew. 2016. “Examining the Evidence Base for Forest Conservation Interventions.” Evidence Gap Map Report, International Initiative for Impact Evaluation, New Delhi.

Rao, Vijayendra, and Michael Woolcock. 2002. “Integrating Qualitative and Quantitative Approaches in Program Evaluation.” In *The Impact of Economic Policies on Poverty and Income Distribution: Evaluation Techniques and Tools*, Francois Bourguignon and Luiz A Pereira da Silva, eds., Chapter 8. New York: Oxford University Press.

Ravallion, Martin. 2016. “How Can We Better Address the Gaps in Our Knowledge about Development Effectiveness?” In *The Oxford Handbook of Professional Economic Ethics*, George DeMartino and Deirdre McCloskey, eds. doi:10.1093/oxfordhb/9780199766635.013.025.

Schwandt, Thomas, Zenda Ofir, Dorothy Lucks, Kassem El-Saddick, and Stefano D’Errico. 2016. “Evaluation: A Crucial Ingredient for SDG Success.” IIED Briefing, International Institute for Environment and Development, London.

Sills, Erin O., Diego Herrera, Justin A. Kirkpatrick, Amintas Brandão Jr, Rebecca Dickson, Simon Hall, Subhrendu Pattanayak, et al. 2015. “Estimating the Impacts of Local Policy Innovation: The Synthetic Control Method Applied to Tropical Deforestation.” *PLOS ONE* 10 (7). doi:10.1371/journal.pone.0132590.

Snilstveit, Birte, Jennifer Stevenson, Daniel Phillips, Martina Vojtkova, Emma Gallagher, Tanja Schmidt, Hannah Jobse, Maisie Geelen, Maria Grazia Pastorello, and John Eyers. 2015. “Interventions for Improving Learning Outcomes and Access to Education in Low- and Middle-Income Countries.” 3ie Systematic Review 24, International Initiative for Impact Evaluation, London.

Snilstveit, Birte, Jennifer Stevenson, Radhika Menon, Daniel Phillips, Emma Gallagher, Maisie Geleen, Hannah Jobse, Tanja Schmidt, and Emmanuel Jimenez. 2016a. “The Impact of Education Programmes on Learning and School Participation in Low- and Middle-Income Countries.” 3ie Systematic Review Summary, International Initiative for Impact Evaluation, London.

Snilstveit, Birte, Jennifer Stevenson, Paul Fenton Villar, John Eyers, Celia Harvey, Steven Panfil, Jyotsna Puri, and Madeleine McKinnon. 2016b. “Land-Use Change and Forestry Programmes: Evidence on the Effects on Greenhouse Gas Emissions and Food Security.” International Initiative for Impact Evaluation, London.

Somanathan, E., R. Prabhakar, and Bhupendra Singh Mehta. 2009. “Decentralization for Cost-Effective Conservation.” *Proceedings of the National Academy of Science of the United States of America* 106 (11): 4143–47. doi:10.1073/pnas.0810049106.

Tachibana, Towa, and Sunit Adhikari. 2009. "Does Community-Based Management Improve Natural Resource Condition? Evidence from the Forests in Nepal." *Land Economics* 85 (1): 107–31. doi:10.3368/le.85.1.107.

UNESCO (United Nations Educational, Scientific and Cultural Organization). 2015. *Education for All 2000–2015: Achievements and Challenges*. Paris: UNESCO.

Waddington, Hugh, and Howard White. 2014. "Farmer Field Schools: From Agricultural Extension to Adult Education." Systematic Review Summary, International Initiative for Impact Evaluation.

Woolcock, Michael. 2009. "Toward a Plurality of Methods in Project Evaluation: A Contextualised Approach to Understanding Impact Trajectories and Efficacy." *Journal of Development Effectiveness* 1 (1): 1–14. doi:10.1080/19439340902727719.

———. 2013. "Using Case Studies to Explore the External Validity of Complex Development Interventions." *Evaluation* 19 (3): 229–48. doi:10.1177/1356389013495210.

World Bank. 2016. *World Bank Annual Report 2016*. Washington, DC: World Bank. doi:10.1596/978-1-4648-0852-4.

Wynes, Seth, and Kimberly A. Nicholas. 2017. "The Climate Mitigation Gap: Education and Government Recommendations Miss the Most Effective Individual Actions." *IOP Science: Environmental Research Letters* 12 (7). doi:10.1088/1748-9326/aa7541.