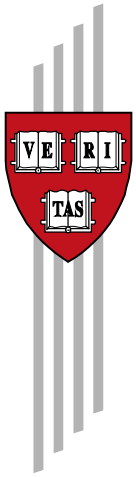
Process and Implementation Evaluations: A Primer

Patricia J. Rogers and Michael Woolcock

CID Faculty Working Paper No. 433
May 2023



Working Papers

Center for International Development at Harvard University

Process and Implementation Evaluations: A Primer

Patricia J. Rogers

Michael Woolcock[1]

May 2023

Abstract

Beyond assessing whether or not interventions have achieved their stated goals, evaluations can also provide greater understanding—in real time and at completion—about how, where, for whom, over what time frame, and which aspects of an intervention may (or may not) have succeeded, and thus where improvements might be sought. Moreover, to the extent any intervention is only as good as its implementation, evaluations can also help identify where breakdowns in the delivery system may have occurred and spaces wherein frontline implementers were able to find innovative solutions to local ("binding constraint") problems. Process and implementation evaluations thus serve the vital purpose of jointly promoting accountability and learning, thereby expanding the common perception of evaluations as external instruments of compliance and discipline to internal procedures for promoting partnerships, innovation, and improvement (organizationally or technically). In this chapter, we describe six different types of process and implementation evaluations and describe their respective strengths and weaknesses in various contexts, for various purposes. As part of collective efforts to enhance the effectiveness of all classes of interventions, impact and process evaluations should be regarded as necessary complements.

[1] The views expressed in this chapter are those of the authors alone and should not be attributed to the organizations (and/or their respective executive directors) with which they are affiliated. Our thanks to Anu Rangarajan and the other editors for their helpful feedback on earlier drafts. Email addresses for correspondence: patricia@betterevaluation.org and mwoolcock@worldbank.org (also michael_woolcock@hks.harvard.edu).

Introduction

Evaluations can serve purposes beyond discerning *whether* a given intervention has met its stated goals (and for whom in what situations), which is the primary task of an impact evaluation. Process evaluations (which are sometimes referred to as implementation evaluations) describe *how* an intervention has worked, exploring details about the program's constituent elements (including the mechanisms connecting them), the combination of elements at particular times, places and people that have driven observed outcomes, and the types of activities that were undertaken—when, where, and for whom.

Almost all evaluations, including process evaluations, are conducted under numerous constraints; rarely is there sufficient time, money, data, support, and staff needed to do what is ideally required—and this is especially true in low-income countries. To optimize under these constraints and generate a process evaluation that is nonetheless useful and credible, it is important to be clear about the intended uses of the process evaluation and to choose methods and procedures best suited to those intended uses and the context in which the evaluation is being conducted.[2] Process evaluations are sometimes assumed to be synonymous with formative evaluations, but this is not entirely accurate. Where process evaluations are distinguished by their focus on identifying and understanding the dynamics of implementation, formative evaluations are distinguished by their purpose—to inform decisions relating to performance improvement. Process evaluations are often intended to be formative, but this is not always or necessarily the case.[3]

The chapter begins with a short overview of the key characteristics of process evaluations. We describe six main types of process evaluations and the particular purposes they serve. We then discuss the array of methods used to conduct process evaluations. We argue that process evaluations are necessary complements to impact evaluations and outline some ways in which more frequent and fruitful partnerships between them might be encouraged. Throughout the chapter, we draw on a range of examples, from developed and developing countries, demonstrating that process evaluations are important in a range of sectors and implementation contexts, especially in the education and health sectors (see, for example, Moore et al. 2015).

1. Overview of Process Evaluation

Some guides on process evaluation equate them with formative evaluation and state that impact evaluation is summative. This approach confuses focus and purpose, which are independent dimensions of an evaluation (Patton 1996). Formative evaluation is intended to support decisions aimed at improvement; summative evaluation is intended to support stop/go decisions, such as which option to choose, which proposal to invest in, and whether or not to continue an intervention (see table 1). While process evaluation is often formative,

[2] A helpful overview of practical strategies for conducting evaluations (of all kinds) under numerous constraints is provided in Bamberger and Mabry (2019).

[3] In the field of international development, Shah et al. (2015) make a helpful distinction between "decision-focused" and "knowledge-focused" evaluations, the latter associated with "rigorous" impact evaluations informing general theory and global knowledge, the former focused primarily on deploying "context-specific tools for decision making that feed into solution-finding systems . . ." (p. ii). In this conceptual space, process evaluations are one such tool for enhancing the empirical and analytical foundations on which policymakers, funders, program managers and frontline implementors make key programmatic decisions, often in real time. As we note below, however, process evaluations can also be part of impact (thus knowledge-focused) evaluations.

sometimes a summative process evaluation is needed. For example, a translation process evaluation (described below) might find that it is not possible to implement an intervention in the ways that have been found to be effective, leading to a decision to not scale it up; a compliance process evaluation (also described below) might find that the quality of implementation is so egregious that the contract with the implementor should be terminated. Similarly, it is possible to have an impact evaluation that is intended to inform decisions about how to improve performance, not whether to continue or end the intervention.

Table 1. *Broad Categories of Evaluation*

	Formative evaluation	**Summative evaluation**
Process evaluation	Focused on processes: intended to inform decisions about improving (primarily) implementation	Focused on processes: intended to inform decisions about stop/go
Impact evaluation	Focused on impact: intended to inform decisions about improving (primarily) design characteristics	Focused on impact: intended to inform decisions about stop/go

Source: Authors' construction

The key questions addressed by process evaluations include questions relating to the following:

The intended beneficiaries. What were the criteria for targeting, prioritizing, inclusion? How were they selected? What was the process for engaging with the intervention? What percentage of the target population was engaged in the program? What worked well, for whom, and in what circumstances in terms of engagement?

The implementing organization. What are the characteristics of the entities or organizations implementing a program? How is the program being implemented?

Services offered. What services were offered? Over what time period? Were these offered in parallel or in sequence? Was there a core package of services or a customized offering? Were there optional services for some participants? Did these services change over time?

Implementation. How does program implementation vary across (and within) implementing organizations, groups, or locations? Who delivered the intervention? What was their existing capacity? How were they selected? How were they prepared for implementation? Was the program delivered as intended (fidelity)? What did not work well in terms of implementation, and why? What are the essential elements needed to make this new intervention work? What are the barriers to implementing and scaling up a successful pilot, and how might these be overcome?

Methods for process and implementation evaluations need to suit the particular purposes of the evaluation. To this end, the chapter describes how to select an appropriate mix of procedures and methods to conduct process and implementation evaluations. It provides examples from a range of countries and sectors but focuses in particular on how such approaches can be undertaken in low-income countries, where time, budget, political, and human resources constraints routinely place severe limits on what can be done, how comprehensively it can be done, and what can be said. These examples elucidate key themes and strategies, showing how challenges can be anticipated and addressed.

In addition to answering questions, the process of doing a process evaluation can also be intended to have impacts—such as infusing evaluative thinking into the organizational culture, enhancing shared understanding between different stakeholders, supporting and reinforcing effective implementation, and increasing civic engagement (Patton 2012).

Process evaluations often benefit by being linked to an intervention's theory of change, which provides an explicit account of the mechanisms by which change is supposed to happen (see Funnell and Rogers 2011). The theory of change can assist identifying important elements to include in data collection and analysis for a process evaluation, and how these might be analyzed; the theory of change can itself be informed by a process evaluation that identifies what are seen to be important activities and what successful implementation looks like on the ground.

2. Different Types of Process Evaluation

While all process evaluations focus on implementation, different types of process evaluation are used for different purposes. In this chapter, we discuss five different types, each with different purposes, drawing on an earlier typology of development evaluation (Rogers and Fraser 2014). Table 2 compares these different types of process evaluation.

Table 2. Types of Process Evaluation

Type	When it is done	Purpose and focus
1. Chronicle	During implementation, before doing impact evaluation (efficacy and effectiveness)	Document what is done (especially in an innovation) to inform future evaluation or scaling up or document an intervention so that decision makers have a better understanding of it
2. Compliance	As part of impact evaluation (efficacy and effectiveness) or rollout/scale-up	Check that actual implementation matches planned implementation as part of an impact evaluation (fidelity) or as part of managing implementation (e.g., quality assurance of the performance of implementing contractors)
3. Translation	After impact evaluation	Identify and overcome barriers to implementation more widely and in other contexts
4. Improvement	As part of rollout	Improve implementation in order to improve results
5. Adaptive Management	In situations of rapid change and uncertainty	Support ongoing learning and adaptation in a complex intervention operating in an environment of unpredictable change and uncertainty

Source: Authors' construction

All of these types of process evaluation can be undertaken by external evaluators, internal staff, or a hybrid team with a mix of internal and external members. Compliance evaluations can be done once or once in a cycle (for example, monthly, annually), and might be undertaken by external evaluators (especially if done as part of an impact evaluation) or by community members or by program staff as part of internal staff orientation and practice quality reviews. For example, community scorecards are often used to record the performance of services such as the opening hours of health services, attendance of teachers,

and the availability of frontline drugs for diseases such as HIV/AIDS, TB, and malaria. Improvement evaluations are particularly likely to involve program staff and/or community members in the evaluation team, potentially gathering and analyzing data and using it to inform ongoing decisions and actions, perhaps with some support from an external expert. We now discuss each of these types of process evaluations in more detail.

2.1. Chronicle—Document and understand implementation of innovations or effective cases

The first type of process evaluation focuses on documenting and understanding how an intervention is being implemented. This is particularly important when trying to document what is being done in innovative projects or particularly successful sites that might be useful for others to learn from. These types of process evaluations can record details of implementation that might otherwise be lost and can inform planning of impact evaluations, especially in terms of identifying what and when to measure. They can also create a vivid, vicarious experience of an intervention that can motivate and inspire others to emulate it or to understand what is needed to support it.

Innovative projects are often poorly documented, which means that even if they are demonstrated to be effective, it can hard to repeat the success—because it is not clear what implementation has involved, what exactly constituted the innovative aspects, how implementers arrived at them, and how authorization to "deviate" from established practice was secured. This is especially important for those interventions that, of necessity, require high levels of discretion and ongoing face-to-face interactions by implementation teams, such as classroom teaching and clinical health care. All teachers, for example, are required to follow a given curriculum designed by education professionals but are also expected to craft lessons tailored to the learning styles and socioeconomic circumstances of their students. A process evaluation can document the array of ways in which the curriculum is interpreted by teaching staff and the extent to which these interpretations give rise to pedagogical efforts that help or hinder the curriculum's realization. This type of process evaluation can be part of a "positive deviance" approach to evaluation, in which successful sites or examples are identified and then studied to see what they are doing differently than other sites (see Pascale, Sternin, and Sternin 2010).[4]

Documenting how innovations have been implemented can be an important step before doing an impact evaluation, as it can help identify what aspects of implementation should be included in data collection and also when outcomes are likely to occur. Documenting implementation and early steps along the impact pathway (that is, early outcomes that occur well before impacts) can also be useful for developing knowledge about the change trajectory of the intervention (Woolcock 2009, 2019a). Most impact evaluations, including randomized controlled trials (RCTs), presume that an intervention's net average effect, as measured by the difference between a baseline and follow-up measure (controlling for known confounding factors), unfolds along a straight line. Such a presumption enables simple projections to be made regarding likely effects in the future. But it does not recognize that the trajectory of

[4] *Positive deviance* refers to the idea that, in complex situations, much can be learned from those who, despite sharing similar characteristics and circumstances to others, nonetheless manage to achieve superior outcomes. How are such outcomes obtained? The idea of mapping, identifying, and understanding positive deviance was originally conceived in public health, where researchers noted that in large, poor urban settlements, children in certain families were far healthier than children in otherwise similar families, due to particular dietary habits adopted by their parents (which in turn could potentially be adopted by neighboring parents, thereby improving child nutrition at greater scale).

change of most interventions, especially those that are complex and/or contextually idiosyncratic—for example, efforts to empower women or to promote the rule of law—are likely to be decidedly nonlinear (and perhaps highly variable across time and space), nor does it account for the reality that the time between baseline and follow-up is largely selected to meet administrative or political imperatives rather than a well-understood theory of change able to specify what impacts it is reasonable to expect by a given time.

For example, an impact trajectory shaped like a "punctuated equilibrium," in which nothing much happens for long periods of time before there is a sudden shift—such as campaigns to end racism or promote marriage equality—would be declared to have had no impact if both the baseline and follow-up data happened to be collected, unwittingly, on the flat part of the impact trajectory curve. Conversely, some interventions may enjoy spectacular initial success (for example, nutrition supplements) and be declared as such following a standard impact evaluation, when deeper understanding of their impact trajectory would recognize that these positive effects are likely to be short-lived.

Chronicle process evaluations are structured to be much more attentive and attuned to such issues, enabling more discerning judgments to be made not just about how interventions work (as discussed above) but by when it is reasonable to expect certain outcomes to be observable for certain groups, in particular places. When the elicited evidence is able to be placed in dialogue with a carefully articulated theory of change sensitive to trajectory issues, it enables evaluators to make more discerning judgements regarding whether outcomes observed at a particular point in time are a function of design characteristics, implementation quality, contextual issues (for example, political interference), selection effects, or (un)reasonable expectations. The answers to these issues are especially important when outcomes might initially appear either disappointing or unusually impressive; if everything else appears to be in order, it can be reassuring (or sobering) to benchmark outcomes against where they should reasonably be at this time, in this context, for this type of intervention—and where they might be after a longer period of time has elapsed. In the case of a seemingly disappointing outcome, it can be comforting to program sponsors and managers to learn that implementing teams may simply need to stick to it for longer; if initial results are unusually positive, it may be instructive to learn that these effects may not endure—or to discover that a genuine innovation has been pioneered that others can replicate. All these key learnings are foreclosed if one draws inferences about effectiveness exclusively on the difference between baseline and follow-up data (even in an RCT), or with little understanding of how, where, and for whom the impact trajectory evolved over time, and how these learnings fared when compared with prior reasoned expectations.

Finally, a chronicle process evaluation that focuses on documenting existing implementation dynamics can also be used to give decision makers a vicarious experience of the intervention or of the experience of those using it, especially where they lack direct experience. For example, the picture shown in figure 1 was used in the Performance Evaluation of the USAID/Malawi Early Grade Reading Activity (USAID 2015). In this report, overcrowding in the classroom was identified as an important context for the new program: "Standard 1 to 3 classes had an average class size of 85, and a range of 5–289 students." The compelling photograph shown in figure 1 was used to convey the level of overcrowding that was involved.

2.2. Compliance—Ensure compliance with intended processes

A compliance process evaluation checks that implementation is complying with what was planned—sometimes with explicit reference to laws, professional norms, program objectives, policies and procedures, budgets, and timelines. It complements other processes that also have a focus on compliance, including quality assurance processes, audits, and some types of monitoring.

A fidelity process evaluation is a particular type of compliance evaluation done as part of an impact evaluation to assess whether and how implementation efforts as planned were in fact conducted. This is essential to be able to identify (and where possible correct) any logistical or behavioral lapses that might cause a program to fail even when it has been well designed and adequately supported, financially and politically. (Such an evaluation might also reveal that an intervention managed to succeed *despite* a weak design because of innovative and persistent efforts by managers and frontline implementers; see Rogers and Macfarlan 2020a, 2020b). There is little point in undertaking an expensive RCT only to conclude that an intervention does not work when it is poorly implemented; what is needed is evidence about how effective it is when implemented well. A compliance process evaluation can both check this and be used to improve implementation either during the trial or during the pilot phase before the trial.

For example, an RCT study of a program seeking to promote participatory democracy in poor rural communities in India was deemed to have had no positive impact (Rao, Ananthpur, and Malik 2017). Upon closer inspection, however, it was learned that there was in fact considerable variation in the program's impact: the average of this variation (that is, the "local average treatment effect") may have been close to zero, but for certain groups the program had worked quite well, while for others it had been detrimental. Who were these groups, and what was it about them that led to such different outcomes? A companion qualitative process evaluation was able to discern that the key difference was the quality of implementation (facilitation) received by different groups, the level of support provided by managers and political leaders, and variations in the nature and extent of local-level inequalities (which in turn shaped which groups were able to participate, and on what terms). The administrative rules and implementation guidelines provided to all groups were identical, but in this case a process evaluation was able to document the ways and places in which variable fidelity to them yielded widely different outcomes. Moreover, the process evaluation was able to discern subtle positive effects from the program that reliance on the quantitative survey instrument alone would have missed.

2.3. Translation—Support translation to new contexts

Among researchers, there are strong incentives to focus on determining *whether* a particular policy or program worked in a particular context. Methodological rigor is highly valued, and when carefully deployed gives (certain) readers greater confidence that a causal relationship has been established between program inputs and outcomes. Such studies are likely to be published in top journals, so researchers prefer to do them. But those overseeing such programs and charged with making key decisions regarding whether they should be replicated elsewhere need vastly more information than a rigorous impact evaluation alone can provide—they also need to know *how* the intervention worked and for whom, how well it was implemented, whether scale and maturity mattered (since replicated programs are likely to be conducted at a scale different than that of their model), and how certain aspects of the

context (for example, its social, economic, and political characteristics) interacted with program design features and implementation dynamics to generate the observed outcomes. These are the key "support factors" (Cartwright and Hardie 2012) that enable outcomes to occur in a particular place (whether or not they are formally observed), and it is their presence or absence in a novel context that will shape whether it is reasonable to presume that an outcome obtained in one context—even those that are the most rigorously verified—can also be expected in another.

Translation process evaluations elicit detailed insights on the form and salience of these support factors, and the ways and places in which they shape variation in the outcomes obtained. An equivalent analysis thus needs to be conducted in the new context where the program is being considered, to discern whether the necessary support factors are present (or if not present, whether credible alternatives can be found).[5] Such considerations are especially salient for interventions that are implementation-intensive (for example, require the active and sustained involvement of local facilitators) and designed to respond to context-specific characteristics, thereby making instantiation of the intervention highly idiosyncratic, and thus inherently likely to generate highly variable outcomes.

An example is the class of projects grouped by the World Bank under Community Driven Development (CDD), which began in Indonesia in the late 1990s and were replicated in (among other places) Cambodia, the Philippines, and Afghanistan. The innovation here was to provide block grants to local communities and establish deliberative councils in compliance with local norms to determine, in an open public forum, which of various proposals submitted to them by small groups of community members were the most promising in terms of feasibility, cost-effectiveness, and broad impact. Among researchers and evaluators, the dominant question in such considerations was, "Does CDD work?" (Mansuri and Rao 2012; Casey 2018), with its implicit presumption that CDD was a singular technology (the social equivalent of a road) that either did or did not work. Given its distinctive characteristics, however, a better and more useful question for replication and translation considerations is, "*When* does CDD work?" (Woolcock 2019b), since the outcomes associated with even the most mature and well-implemented CDD programs are inherently heterogenous. This is especially so when they are assessed on their ambitious goal of not just improving incomes but enhancing the quality of local governance—because of their deep dependence on skilled facilitation, the direct challenges program rules often pose to prevailing power structures (overseen by leaders who may overtly resist, for example, program requirements for transparency, accountability, and gender equality), and the potential for conflict they generate (by creating winners and losers in the competition for finite funds). It helps to know that CDD has demonstrably worked in highly diverse contexts such as Indonesia (Barron, Diprose, and Woolcock 2011) and in politically fragile contexts such as Afghanistan (Beath, Christia, and Enikolopov 2015), but decision makers considering whether to adopt it in (say) Cameroon need to know much more about the particular scope of conditions under which (or support factors by which) it has and has not worked.

2.4. Improvement—Enhancing quality of implementation

Even if a program is entirely in compliance with prevailing laws, professional norms, program rules, and the like, it may still be falling short of its objectives; conceivably, it may be vastly exceeding them. In either case, it is important to understand how, where, why, and

[5] The spirit of searching for salient support factors also includes identifying those factors that might be inhibiting more positive outcomes from being obtained or compromising their impact on particular groups.

for whom these outcomes are being obtained. Where translation evaluations, outlined above, focus on identifying the broader support factors salient for considering whether a given intervention should be scaled or replicated, improvement process evaluations focus on unpacking the constituent mechanisms articulated in the intervention's theory of change, examining in detail the particular links in the chain where either breakdown or innovation is occurring during implementation. (There can be literally thousands of such links and thus corresponding decision points where implementation can go right and/or wrong.)

In principle, monitoring procedures can perform some of these tasks, but such procedures are usually administrative in form and function, and so may not yield insights of sufficient granularity. For example, if unanticipated problems arise in unanticipated ways in unanticipated places, or if innovative solutions are being deployed during implementation, it is unlikely (by definition) that these will be detected by prevailing administrative instruments. Improvement process evaluations can help elicit both these problems and their solutions, and thereby provide more detailed understandings of where and how improvements might be sought.

It is not unreasonable to suggest that had improvement evaluations been introduced earlier and more systematically into the implementation of Peru's One Laptop Per Child program, they would surely have enhanced the effectiveness of the $225 million spent by the Government of Peru on this otherwise well-intentioned effort—which was designed by experts from MIT and actively championed by the United Nations Development Programme but yielded no significant changes in its core objective, namely to improve student learning outcomes.[6] Subsequent assessments found that the program gave vastly inadequate attention to (among other things) teacher training in the use of laptops (generally, and for classroom teaching in particular) and the logistics of how basic repair/maintenance issues would be conducted.

The information yielded by an improvement evaluation can also be useful when evaluators cannot get long-term data but are called upon to inform changes along the way. This can be especially important when evaluators need to respond professionally in situations where political imperatives demand early, positive results, but such results cannot reasonably be provided or substantiated (for example, when an intervention assumes the status of being a high-profile politician's signature program that thus becomes "too important to fail", especially if an election is imminent). Improvement evaluations can contribute constructively in such situations by speaking directly to specific aspects or places where clear successes or challenges are emerging. Identifying and explaining local variations in response to COVID masking-wearing policies and vaccine rollout strategies are one contemporary example. At the height of the pandemic, the political imperative was very strong to declare victory and attribute this to a particular policy or programmatic intervention. Such was the complexity and uncertainty at this point surrounding the efficacy of *any* specific response, however, let alone the effects of their interaction with one another and of local contextual factors, that the optimal contribution would have been findings from an improvement evaluation documenting the effectiveness of particular constituent mechanisms—that is, of social distancing, contact tracing, mask wearing, and so on.

A final contribution that improvement evaluations can make is discerning whether more intensive action research or some other way of gathering evidence is needed to identify opportunities for performance improvement. This may include making some level of causal inference (that is, of discerning which principles, people, processes, and practices—and not

[6] See the findings of RCT impact evaluations reported in Beuermann et al. (2015) and Cristia et al. (2017).

others—are primarily driving particular outcomes in particular ways for particular people). Such discernment requires a close dialogue between collected evidence and the intervention's theory of change—and thus formally articulating such a theory if one is not explicitly provided, since it is crucial to understand the specific causal mechanisms and pathways by which an intervention's inputs connect to its stated outcomes, and the extent to which these are actually salient in practice. This action research can also include incorporating small-scale rapid cycle or A/B tests within particular aspects of an intervention, the better to help managers make informed choices about discrete technical issues (for example, testing different methods to reduce no-shows at medical appointments via text reminders).[7]

2.5. Adaptive Management—Support ongoing innovation and adaptation

Adaptive Management process evaluations are undertaken as a key input to building implementation capability in teams tasked with responding to policy challenges characterized by deep uncertainty—that is, those wherein it is inherently difficult to cleanly prespecify all aspects of the problem and solution, thus requiring extensive, continuous, real-time innovation by managers and frontline implementers. In effect, such initiatives—for example, Problem-Driven Iterative Adaptation (PDIA)[8] and related approaches—merge evaluation and implementation, incorporating feedback from a real-time process evaluation into the change procedures themselves, the better to foster continuous learning and adjustment, and to assess the extent to which problems are being accurately identified and (hopefully) solved.

An instructive example of an Adaptive Management process evaluation (fused with efforts to enhance effective implementation) comes from Sri Lanka, where a team deploying the PDIA approach was asked by the Ministry of Finance to enhance its capacity to attract direct foreign investment. For its level of development, and in comparison with neighboring countries, Sri Lanka was deemed to be considerably underperforming with respect to its ability to attract foreign investment; yet if the broad problem was clear, the precise factors needing to be changed were not, and thus neither were the solutions. Most "experts" consulted on this question provided what they viewed as technical solutions—improve the regulatory environment, cut red tape, provide better tax incentives, and so on—when the actual underlying problem, only identified after many weeks of analysis, was that staff in the direct foreign investment unit actually had little understanding of how past or potential investors experienced Sri Lanka and didn't know how to get that information. Going through the PDIA process enabled staff to nominate and prioritize their problems, to find their own solutions to them, and to craft a credible strategy for implementing them—which entailed, among other things, compiling their country's first database on past and current investors, including those who has made initial inquiries but dropped out. Contacting all these investors, which took months of patience, persistence, and refinement, was far from a glamorous or easy technical fix, but by doing it themselves (as opposed to outsourcing the task to consultants), the team acquired a vastly more granular understanding of their world and ultimately secured millions of dollars in foreign investment in solar technologies (Andrews et al. 2017).

[7] See Berliner Senderey et al. (2020) for a more detailed discussion and example of deploying of rapid cycle evaluations and A/B tests in evaluation.

[8] PDIA is a practical strategy for enhancing the capability of organizations (particularly those in the public sector) to implement their policies and programs. See Andrews, Pritchett, and Woolcock (2017).

3. Methods Used in Process Evaluation

3.1. Overview of methods for different types of questions

The methods that will be appropriate for data collection and analysis in a process evaluation depend on three factors, including the nature of the evaluation (especially the type of process evaluation, as discussed above, the types of questions being asked, and the information preferences of the primary intended users); the nature of what is being evaluated (especially how visible the activities and associated changes are); and the available resources and constraints (especially time, money, expertise, existing data, and ability to travel, which can be curtailed by conflict, disease, and extreme weather).

A good theory of change that draws on evidence can be helpful in guiding data collection for answering all types of questions, including descriptive questions. It can identify the types of conditions that are important to examine, what characteristics of activities need to be considered, and what behaviors are important (Funnell and Rogers 2011).

Clarifying the type of process evaluation needed, and therefore the questions being asked, can guide data collection, analysis, and reporting.

It is also helpful to consider the different types of questions (descriptive, causal, and evaluative) in an evaluation, as they need different methods and designs.
For example, consider a process evaluation of a primary health service's efforts to improve the engagement of clients from marginalized groups in the community. The overall question might be, "How effective have been efforts to increase engagement from marginalized groups?" To answer this, the evaluation will need to answer the three different types of questions listed in table 3.

Table 3. Disaggregating a Key Evaluation Question into Different Types of Questions

Key Evaluation Question: How effective have been efforts to increase engagement of clients from marginalized groups?	
Descriptive question	What activities were undertaken to try to increase engagement? In what ways did these vary across sites and staff? What were the reactions of the community to these efforts? How has the level of engagement changed? *For example, are there now higher numbers of people from marginalized groups attending the health service, or are they more actively engaging in activities?*
Causal question	Why did observed changes happen? *If engagement levels have increased, have the activities of the health service to increase engagement actually contributed to increased levels of engagement—or has this been due to other factors such as improved local transport that was not part of the improvement project?*
Evaluative question	How good is that? *If engagement levels have increased, is this good enough? For example, is a 10% increase over three months enough to be counted as a success? Is it enough for the marginalized group to now be accessing the service at the same rate as the average—or should it be even higher given their higher rate of health needs?*

Source: Authors' construction

The following sections discuss each of these in more detail.

3.2. Methods for answering descriptive questions about what has happened

Descriptive questions in a process evaluation ask about how things are and what has happened. This can include describing social, economic and environmental conditions before and after (and during) a project or program, describing the activities done to implement a project or program, and describing the behaviors of those involved. A wide range of possible methods can be used to collect data to answer descriptive questions, including direct observation (in real time or through video or photos, including satellite data), interviewing people about their observations or their recollections of their own actions, and drawing on official records of implementation.

Table 4 shows some possible data sources and when these might be collected, in terms of some possible questions and sub-questions that might be asked in a process evaluation.

Table 4. Possible Data Sources and Timing for Answering Descriptive Questions

Possible questions and sub-questions	Possible data sources	When data might be collected
What is the level and nature of engagement in the program by people from marginalized groups? In what ways does this vary across sites and staff?	* Program records of enquiries, enrolments, service delivery, attendance * Key informant interviews from program staff, community members * Direct observation of behavior * Self-report by participants through a survey, individual interview, or group interview	* Before intervention (either directly or using existing data) * During intervention (tracking any changes) * After intervention (including following up to see if changes are maintained)
What activities were undertaken to increase engagement?	* Document review of artifacts such as flyers and social media posts * Direct observation * Interviews with staff, intended participants	* During intervention * After intervention
What were the reactions to these activities?	* Analytics of social media engagement * Interviews with staff, intended participants * Survey of intended participants	* During intervention * After intervention

Source: Authors' construction

For example, if evaluating efforts to increase engagement by people from marginalized communities, it might be possible to draw on the records of the health service—if these are accurate, comprehensive, and accessible (in practical terms and in terms of ethical considerations). However, not all services accurately record demographic details of those who attend. It might be therefore better to directly observe who attends, or to interview people about their attendance, or to speak with key informants who are seen to be able to provide a credible assessment of attendance. For some types of activities, it might be possible to draw on other types of data. For example, wear patterns on equipment and flooring can show where there is high foot traffic. Mobile phone data can show changes in travel.

Another important aspect of answering descriptive questions is choosing an appropriate sampling method from the three broad types—volunteer/convenience sampling, random sampling, and purposeful sampling. Volunteer and convenience sampling involve gathering data from readily accessible sites or people, and risks producing misleading results that don't reflect the population. Random sampling supports statistical generalization but requires an adequate sampling frame, response rate, and sample size. It is not possible to statistically generalize from a sample which is systematically skewed, for example by a very low response rate or lack of coverage of hard-to-reach communities. Purposive sampling selects information-rich cases from a given population to make analytical inferences about the population. For example, extreme case sampling might be used to argue that if implementation is not being done well even in well-resourced sites where there is the intention to implement it, then it is likely that implementation is even worse in other sites. Maximum variation sampling or theoretical sampling can more efficiently test the robustness of findings and consistency with the theoretical models underpinning the intervention. It can be important to collect data from people located at different points on key variables presumed to shape outcomes (for example, gender, age, occupation, income) and conducting enough in-depth interviews to see if patterns are consistent across these, or if they vary. (For further information on purposeful sampling, see Patton [2015], and for a visual overview see Vaca [2018]).

Ensuring an adequate response rate is important when using surveys. For example, if less than 50 percent of participants reply to a questionnaire, it is likely that those who do not respond will differ systematically in some ways—they are more likely to be less engaged in the service or marginalized in some ways.

3.3. Methods and designs for answering causal questions

It is sometimes incorrectly assumed that only impact evaluations address causal questions. However, process evaluations often need to not only describe what has happened at the level of activities but also explain why. For example, if a service is trying to increase the level of participation by people from marginalized groups and finds that the level has indeed increased, it will be important to understand why this has occurred, and in particular which aspects of outreach and improved accessibility have made a difference and should be continued (see chapters on performance evaluations in this volume).

In most cases, there are multiple contributing factors to observed changes, and an effective process evaluation will identify potential factors and use appropriate causal inference strategies to investigate them.

When trying to improve implementation processes, sometimes it is important to address a number of factors at the same time, and sometimes it is sufficient to address them one at a time. For example, if a program is trying to increase participation in a service, different people might have different barriers to engagement—some people might know that the service exists, some might have transportation to be able to attend a service, and some might not have sufficient trust in the service providers. If there is only one barrier to engagement, then overcoming it can improve participation—for example, trialing a transport service would make a difference for those who know about the service and want to attend but cannot get there. It is a different situation where more than one barrier exists. For example, if an organization is trying to change staff practices, there might need to be attention to improving capacity, motivation, and opportunity (drawing on the behavior wheel theory of change from

Michie et al. 2011). Simply providing training will not be effective in terms of changing practice, even if it succeeds in improving skills and knowledge, if there are still barriers in terms of motivation (the new practice is not seen as appropriate by staff or encouraged by management) or opportunity (lack of equipment or scheduling constraints means it cannot practically be implemented).

We describe some commonly used approaches in process evaluations that are used to address causal questions.

Contribution analysis (see Mayne 2001, 2019) is an approach to causal inference that explicitly recognizes the other factors that combine with most interventions to produce the intended effects. It assesses whether an intervention made a difference in terms of it being a necessary part of a package of causes that together brought about or contributed to the important changes. In addition to answering the question of whether an intervention made a difference, it also answers the question, "How and why has the intervention (or component) made a difference, or not, and for whom?"

A process evaluation can draw on different approaches to causal inference. A number of these are discussed in Rangarajan (forthcoming), including rapid cycle evaluations and A/B testing. Two other approaches, process tracing and Qualitative Impact Protocol (QuIP), are discussed briefly below.

Process tracing (Punton and Welle 2015) uses detailed, within-case empirical analysis of how a casual process plays out in a real case. For example, a service seeking to improve attendance in appointments, such as routine vaccines or regular health checkups, might have used the opportunity of a community festival to promote the service and encourage attendance. Process tracing could be used to investigate clients who attended after this to see if their attendance was consistent with being influenced by the promotion. New clients who attended after the festival might be asked if they had been to the festival and remembered the promotion or had heard about if from someone else. They might be directly asked what prompted them to attend the service. Process tracing would pay particular attention to the "negative cases"—new clients who had not been influenced directly or indirectly by the festival promotion, and festival attendees who did not engage with the health service—to understand the likely scale of change, and the potential impact of other factors.

Participant attribution is another approach that relies on testimony from participants about the importance of different factors in explaining changes in observed behavior. Participant attribution presents many obvious challenges in terms of validity—in particular, people are not always aware of the relative importance of different factors that have influenced their behavior, and their answers might be influenced by social desirability and by a desire for a positive evaluation of a local service. QuIP is an important methodological innovation that has been designed to address these concerns while providing a practical approach for small evaluations (Copestake, Morsink, and Remnant 2019; ART 2018; Bath Social and Development Research 2018; Wright and Copestake 2004). QuIP elicits from participants those changed aspects of their lives to which they credit an intervention. This involves conducting interviews with program participants using local researchers, but with the interviewers themselves being "blinded"—that is, they have not been made privy to the name or nature of the specific development intervention that they are asking about. Instead, respondents are asked about social, political, and economic changes that may have taken place in their lives, the factors to which they attribute these changes, and the particular ways in which these factors may have brought about these changes. In effect, respondents are asked

to work backward from outcomes to the decisive combinations of processes and resources that contributed to them. Such analyses have been done in a variety of countries and sectors and might be particularly relevant in the field of international development and in contexts where small-scale efforts undertaken by community organizations, small businesses, and local governments lack the scale and resources to undertake good-quality counterfactual, regularity or theory-testing causal designs.

3.4. Methods for answering evaluative questions about how good something is

Answering evaluative questions requires a combination of explicit values about what is seen as good or bad (or better or worse) and evidence about how things are (which is gathered through answering descriptive questions). It can be helpful to think explicitly about evaluative questions in terms of four components:

1. Criteria: What dimensions of performance are relevant—these could relate to client behavior, such as using services; staff behavior, such as treating clients respectfully; or organizational capacity, such as opening hours or availability of equipment
2. Standards: What levels of performance are desirable—for example, does using services mean attending health services at all, or at a level consistent with recommended attendance, including keeping to the vaccination schedule
3. Evidence: What evidence will be collected and how
4. Synthesis: How evidence about performance across a number of dimensions will be combined—especially in terms of whether or not some elements are considered essential or can be balanced out by performance on other dimensions

Sometimes criteria and standards are formally stated in policies, research evidence, and commitments, such as in the Sustainable Development Goals. For example, SDG 3.8 states as a process goal the achievement of "universal health coverage, including financial risk protection, access to quality essential health-care services and access to safe, effective, quality and affordable essential medicines and vaccines for all."

Sometimes the values that should be used to answer evaluative questions are not adequately documented or agreed on, and in some cases might be tacit and not documented at all. In these cases, to articulate and document tacit values, methods and processes such as the following are needed:[9]

- Hierarchical card sorting—where individuals rank cards representing different cases in terms of their level of success, clarifying their tacit values
- Photovoice—where individuals take photographs of what is important in a community and share the photographs and their explanations to support discussions to surface tacit values
- Rich pictures—where a group explores, acknowledges, and defines a situation and expresses it through diagrams to open discussion and come to a broad, shared understanding of a situation
- Values clarification interviews or questionnaires—which collect data from individuals about what they value
- Most Significant Change approach—which involves groups voting on what they rate as the most significant stories of change

[9] Information on all these methods, processes, and approaches can be found at www.betterevaluation.org.

If there is not agreement about the values that should be used, then processes are needed to negotiate between these different values, such as dotmocracy, public consultations, and critical systems heuristics.

Three broad approaches to answering evaluative questions in a process evaluation can be identified:

- Quality control—this involves checking compliance with a well-established level of performance
- Expert judgement—this involves an expert providing a rating based on their expert knowledge of what constitutes good performance and how it might be assessed
- Rubric—developing and using one or several global rating scales that provide for synthesis of diverse evidence in terms of explicit criteria and standards

A quality control approach to answering evaluative questions depends on having clearly developed and justified criteria, standards, and evidence types. For example, it would be appropriate to observe whether or not a service was open during advertised times, or whether rooms were overcrowded. Sometimes it is appropriate for data collection to be done internally by program staff, or as a reciprocal peer evaluation between staff or sites, and sometimes it is important to have an external data collector, or robust automatic systems, to ensure data validity and credibility. In such cases it is also important to ensure the integrity of the data through ensuring independence of the data collector, adequate training to ensure consistent data collection, and ensuring that the indicators adequately represent the range of performance and cannot be gamed or corrupted. For example, if "prompt answering of telephone inquiries" was a criterion, and the standard was "all phone calls answered within 1 minute," it would be important to ensure that reported performance was not being misrepresented through a system where many incoming calls were disconnected before being answered or were redirected in ways that started the timing again.

Expert judgment can be an important part of a process evaluation and is often done as part of a field visit. Care is needed to ensure that the criteria, standards, and evidence being used are valid. In particular, there can be risks in relying uncritically on the judgments of experts who might be unaware of the effect of cognitive biases, such as people's tendency to seek data that will confirm their initial judgements (Kahneman 2011).

Rubrics, sometimes referred to as global rating scales, provide a rating that is based on transparent and defensible evidence and values based on agreed criteria, standards, and synthesis. Rubrics also support processes of articulating differing values among key stakeholders. Rubric development involves identifying dimensions of performance, developing descriptions of what performance at different levels would look like (for example, what would be considered very good, OK, or not OK), and then identifying how evidence can be gathered and synthesized to produce an overall rating. This overall rating is usually not a simple averaging of data. For example, "engagement with clients" might be rated as "not OK" if there is evidence that some clients feel disrespected, even if other aspects of this have evidence of good performance. An example of using rubrics to evaluate a program to support first-time school principals can be found in UNICEF's methodological brief on Evaluative Reasoning (Davidson 2014).

The following example shows how a range of evidence and processes might be combined to answer evaluative questions in a process evaluation. The community accountability project Vidya Chaitanyam supported women in existing women's self-help groups in rural Andhra Pradesh, India, to undertake regular evaluations of school quality in a simple scorecard and to publish the results in group meetings and at school management committees in order to pressure to their local primary schools to improve quality (Galab et al. 2013).

The scorecard covered five dimensions of school quality, most relating to process and some relating to outcomes: (1) student progress (attendance, academic performance), (2) leadership and management (display of school statistics, use of government grants, parental support), (3) teaching and learning (teacher attendance, use of teaching and learning materials, cocurricular activities), (4) infrastructure (maintenance, provision of midday meal, toilets), and (5) image of the school (parental involvement in school management meetings).

Care was taken to ensure the credibility of the scorecard and of the data collected using it. The indicators were designed in consultation with district-level officials and in line with indicators already being used at the state level. The cutoff points (between the three levels in the traffic light rating scale) were chosen to align with existing indicators, and standards were agreed on collectively. To ensure that data could be reliably and validly collected by the community members (who had a 70 percent illiteracy rate), the scorecard used clear symbols. The project included training and supporting marginalized women in the community, including those who were illiterate and semiliterate, to gather, record, discuss, and present the scorecard assessments. (For further information on community scorecards and their use in process evaluations, see CARE Malawi 2013 and Kosack et al. 2021.)

4. Conclusion

Process evaluation is important throughout the program cycle—from early stages of developing an innovative approach, to contributing to proof of concept, efficacy and effectiveness evaluations, supporting decisions about cancelation, replication and scaling up, and ongoing adaptation and improvement. Process evaluations can enhance the quality of implementation, provide explanations of how both average and particular impacts were attained, and inform decisions about the likely conditions under which similar findings might be expected elsewhere or when implemented at scale. Effective process evaluations can inform decisions about legal, administrative, financial, and political aspects of policies and programs, and in doing so can help enhance the welfare of millions of people.

As such, process evaluations should be a key tool in every evaluator's kit and a part of planning the evaluation of all interventions. To do them well, however, requires integrating skill sets from across disciplines, familiarity with the respective strengths and weaknesses of particular methodologies, and the confidence to function professionally under an array of shifting constraints (Stern et al. 2012).

References

Andrews, Matt, Duminda Ariyasinghe, Krishantha Britto, Peter Harrington, Nelson Kumaratunga, M. K. D. Lawrance, Tim McNaught, et al. 2017. "Learning to Engage New Investors for Economic Diversification: PDIA in Action in Sri Lanka." Cambridge: Harvard University, Center for International Development Faculty Working Paper No. 336.

Andrews, Matt, Lant Pritchett, and Michael Woolcock. 2017. *Building State Capability: Evidence, Analysis, Action*. New York: Oxford University Press.

ART (Assessing Rural Transformations). 2018. "Qualitative Impact Protocol (QuIP): Guidelines for Field Use." https://reshare.ukdataservice.ac.uk/852065/2/ProjectDescription_Guidelines.pdf.

Bamberger, Michael, and Linda Mabry. 2019. *RealWorld Evaluation: Working Under Budget, Time, Data, and Political Constraints*. 3rd ed. Thousand Oaks, CA: SAGE Publications.

Barron, Patrick, Rachael Diprose, and Michael Woolcock. 2011. *Contesting Development: Participatory Projects and Local Conflict Dynamics in Indonesia*. New Haven, CT: Yale University Press.

Bath Social & Development Research. 2018. "Comparing QuIP with 30 Other Evaluation Approaches." http://bathsdr.org/wp-content/uploads/2018/04/Comparing-QuIP-with-thirty-other-approaches-to-evaluation.pdf.

Beath, Andrew, Fotini Christia, and Ruben Enikolopov. 2015. "The National Solidarity Programme: Assessing the Effects of Community-Driven Development in Afghanistan." *International Peacekeeping*, 22 (4): 302–320. doi: 10.1080/13533312.2015.1059287.

Berliner Senderey, Adi, Tamar Kornitzer, Gabriella Lawrence, Hilla Zysman, Yael Hallak, Dan Ariely, and Ran Balicer. 2020. "It's How You Say It: Systematic A/B Testing of Digital Messaging Cut Hospital No-Show Rates." *PLOS ONE*, 15 (6): e0234817.

Beuermann, Diether W., Julian Cristia, Santiago Cueto, Ofer Malamud, and Yyannu Cruz-Aguayo. 2015. "One Laptop per Child at Home: Short-Term Impacts from a Randomized Experiment in Peru." *American Economic Journal: Applied Economics*, 7 (2): 53–80.

CARE Malawi. 2013. CARE Malawi. "The Community Score Card (CSC): A Generic Guide for Implementing CARE's CSC Process to Improve Quality of Services." Cooperative for Assistance and Relief Everywhere, Inc.

Cartwright, Nancy and Jeremy Hardie. 2012. *Evidence-Based Policy: A Practical Guide to Doing it Better*. New York: Oxford University Press.

Casey, Katherine. 2018. "Radical Decentralization: Does Community-Driven Development Work?" *Annual Review of Economics*, 10: 139–163.

Copestake, James, Marlies Morsink, and Fiona Remnant (eds). 2019. *Attributing Development Impact: The Qualitative Impact Protocol Case Book*. Rugby, UK: Practical Action Publishing.

Cristia, Julian, Pablo Ibarrarán, Santiago Cueto, Ana Santiago, and Eugenio Severín. 2017. "Technology and Child Development: Evidence from the One Laptop per Child Program." *American Economic Journal: Applied Economics*, 9 (3): 295–320.

Davidson, E. J. 2014. "Evaluative Reasoning." UNICEF methodological brief. Office of Research, Florence. https://www.unicef-irc.org/publications/pdf/brief_4_evaluativereasoning_eng.pdf.

Funnell, Sue, and Patricia Rogers. 2011. *Purposeful Program Theory: Effective Use of Theories of Change and Logic Models*. New York: John Wiley & Sons.

Galab, S., Charlotte Jones, Michael Latham, and Richard Churches. 2013. "Community-Based Accountability for School Improvement: A Case Study from Rural India." Reading, UK: CfBT Education Trust. https://www.educationdevelopmenttrust.com/EducationDevelopmentTrust/files/2d/2d97cea6-93f3-42d8-93db-f1ece9c7953b.pdf.

Kahneman, Daniel. 2011. *Thinking Fast and Slow*. New York: Farrar, Straus and Giroux.

Kosack, Stephen, Jessica Creighton, Courtney Tolmie, Fatu Conteh, Eric Englin, Linda Gassama, Hannah Hilligoss, et al. 2021. "Brokering Collaboration: Involving Officials in Community Scorecard Programs." Ash Center Occasional Paper Series. Cambridge, MA: Harvard Kennedy School.

Mansuri, Ghazala, and Vijayendra Rao. 2012. *Decentralizing Development: Does Participation Work?* Washington, DC: World Bank.

Mayne, John. 2001. "Addressing Attribution Through Contribution Analysis: Using Performance Measures Sensibly." *Canadian Journal of Program Evaluation*, 16 (1): 1–24.

Mayne, John. 2019. "Revisiting Contribution Analysis." *Canadian Journal of Program Evaluation*, 34 (2): 171–191.

Michie, Susan, Maartje M. Van Stralen, and Robert West. 2011. "The Behaviour Change Wheel: a New Method for Characterising and Designing Behaviour Change Interventions." *Implementation Science*, 6 (1): 1–12.

Moore, Graham F., Suzanne Audrey, Mary Barker, Lyndal Bond, Chris Bonell, Wendy Hardeman, Laurence Moore, et al. 2015. "Process Evaluation of Complex Interventions: Medical Research Council Guidance." *British Medical Journal*, 350: h1258. https://doi:10.1136/bmj.h1258.

Pascale, Richard, Jerry Sternin, and Monique Sternin. 2010. *The Power of Positive Deviance*. Boston: Harvard Business Press.

Patton, Michael Quinn. 1996. "A World Larger Than Formative and Summative." *Evaluation Practice*, 17 (2): 131–144.

Patton, Michael Quinn. 2012. *Essentials of Utilization-Focused Evaluation*. Thousand Hills, CA: SAGE Publications.

Patton, Michael Quinn. 2015. *Qualitative Research and Evaluation: Integrating Theory and Practice.* 4th ed. Thousand Hills, CA: SAGE Publications.

Punton, Melanie, and Katherina Welle. 2015. "Straws-in-the-Wind, Hoops and Smoking Guns: What Can Process Tracing Offer to Impact Evaluation?" CDI Practice Paper 10. Brighton, UK: IDS.

Rangarajan, Anu (ed.) (forthcoming) *Oxford Handbook of Program Design and Implementation* New York: Oxford University Press, pp. 294-316.

Rao, Vijayendra, Kripa Ananthpur, and Kabir Malik. 2017. "The Anatomy of Failure: An Ethnography of a Randomized Trial to Deepen Democracy in Rural India." *World Development*, 99 (11): 481–497.

Rogers, Patricia, and Alice Macfarlan. 2020a. "An Overview of Monitoring and Evaluation for Adaptive Management." BetterEvaluation Working Paper 1. https://www.betterevaluation.org/resources/overview-monitoring-and-evaluation-adaptive-management-working-paper-1.

Rogers, Patricia, and Alice Macfarlan. 2020b. "What Is Adaptive Management and How Does it Work?" BetterEvaluation Working Paper 2. https://www.betterevaluation.org/resources/what-adaptive-management-and-how-does-it-work-working-paper-2.

Rogers, Patricia, and Dugan Fraser. 2014. "Development Evaluation." In *International Development: Ideas, Experience, and Prospects*, edited by Bruce Currie-Alder, Ravi Kanbur, David M. Malone, and Rohinton Medhora. New York: Oxford University Press.

Shah, Neil Buddy, Paul Wang, Andrew Fraker, and Daniel Gastfriend. 2015. "Evaluations with Impact: Decision-Focused Impact Evaluation as a Practical Policymaking Tool" 3ie Working Paper No. 25, New Delhi: International Initiative for Impact Evaluation.

Stern, Elliot, Nicoletta Stame, John Mayne, Kim Forss, Rick Davies, and Barbara Befani. 2012. "Broadening the Range of Designs and Methods for Impact Evaluation." London: UK Government, Department for International Development, Working Paper 38 . https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/67427/design-method-impact-eval.pdf .

USAID. 2015. "Performance Evaluation of the USAID/Malawi Early Grade Reading Activity (EGRA)." Washington, DC: USAID. https://pdf.usaid.gov/pdf_docs/PA00KBNS.pdf.

Vaca, Sara. 2018. *Patton's 40 Purposeful Sampling Strategies.* https://www.saravaca.com/project/pattons-40-purposeful-sampling-strategies/.

Woolcock, Michael. 2009. “Toward a Plurality of Methods in Project Evaluation: A Contextualized Approach to Understanding Impact Trajectories and Efficacy.” *Journal of Development Effectiveness*, 1 (1): 1–14.

Woolcock, Michael. 2019a. “Reasons for Using Mixed Methods in the Evaluation of Complex Projects.” In *Contemporary Philosophy and Social Science: An Interdisciplinary Dialogue*, edited by Michiru Nagatsu and Attilia Ruzzene, pp. 149–171. London: Bloomsbury Academic.

Woolcock, Michael. 2019b. “When Do Development Projects Enhance Community Well-Being?” *International Journal of Community Well-Being*, 2 (2): 81–89.

Wright, Katie, and James Copestake. 2004. “Impact Assessment of Microfinance Using Qualitative Data: Communicating Between Social Scientists and Practitioners Using the QUIP.” *Journal of International Development*, 16 (3): 355–367.