Development Evaluation

Patricia J. Rogers and Dugan I. Fraser

Abstract

Evaluation has a long history of being used to promote and improve development. Different evaluation approaches have been used, reflecting different understandings of what development is and the role of evaluation—as a support for rational decision making or as part of an incentive system for performance. These different approaches to development evaluation are not mutually exclusive or incommensurable. Each has potential validity and utility, depending on what is being evaluated and the purpose of the evaluation. We discuss eight approaches to development evaluation, noting that most have been present throughout the history of development but have been emphasized and promoted at different historical points for political and other reasons. Evaluators and development practitioners who understand the full range of evaluation options and the types of development for which they
are best suited can maximize the contribution evaluation can make to improving development processes and outcomes.

**Keywords:** evaluation, performance, accountability, learning, improvement, evidence
Introduction

Evaluation has long been considered an integral part of effective development, but over time there have been different views about how development evaluation should be undertaken. This chapter shows how these different approaches to development evaluation reflect different ideas about what development is and how evaluation can contribute to improving it.

If development is primarily understood to be about implementing effective development projects, then development evaluation needs to support this by helping to choose the right projects to invest in, and supporting their management. If development is primarily understood to involve scaling up effective technologies within a country or across countries, then development evaluation needs to focus on identifying effective technologies, communicating these findings clearly to policy makers, and monitoring implementation fidelity during scaling up. If development is primarily about building the capacities of government and non-government organizations to create, adapt, and implement technologies that are appropriate to their diverse and changing situations, then development evaluation needs to build capacity, supporting implementers to improve their practice. (Note that these different assumptions reflect the distinctions that Ocampo (in this volume) has drawn between different approaches to “technological catching up.”)
All of these concepts focus on evaluation’s role in informing rational decision making, but there are other ways of thinking about how evaluation can influence development. Evaluation can contribute to positive incentives for good performance and to negative incentives (sanctions) for poor performance through its role in accountability to funders or to the community. By reassuring funders that money is being well spent, evaluation can also serve an important function in ensuring that effective programs and organizations receive ongoing funding.

Rather than seeing these as mutually exclusive, incommensurable approaches to development evaluation, in this chapter we discuss them as potentially appropriate choices for particular situations, depending on what is being evaluated and the purpose of the evaluation. None of the approaches described here are irrelevant, nor are any of them universally appropriate.

Despite the somewhat chronological presentation of different approaches in this chapter, there has not been a simple progression of theory and practice: in fact the various different approaches highlighted sequentially in this chapter have for the most part been present throughout the history of development evaluation. Specific techniques have simply been emphasized and promoted more successfully than others as a result of the interplay of a range of different political, economic, and ideological factors at specific conjunctures.
In part this is because the different approaches retain some degree of relevance and appropriateness, at least in some circumstances. In part the lack of informed evolution of evaluation theory and practice is due to the limited knowledge of earlier evaluation theory and practice among practicing evaluators and those who develop evaluation guidelines. Because in most contexts there is no accreditation or certification among evaluators, no agreed curriculum or required training, and because evaluation draws on many different disciplines that tend to publish in different journals and books, there is often little understanding of what has already been done and what is currently being done in different disciplinary or organizational circles. This chapter has attempted to bring together insights from different disciplines and sectors that often co-exist in mutual ignorance.

**Informing investment decisions by donors**

Early development evaluation began at the same time as the growth in international aid following the Second World War. In a review of evaluation across different time periods, Basil Cracknell, former Chair of the OECD’s Expert Group on Evaluation, pointed to the emphasis in the late 1960s and 1970s on “economic project appraisal, which was seen as crucial to good project selection and formulation” (Cracknell 2000). Little and Mirrlees
(1968, 1974), initially in an OECD manual and then in revised form in a book, set out a process for ex ante evaluation of proposed projects, including the use of shadow prices for non-traded goods.

Ex ante evaluation, sometimes called economic project appraisal, is intended to improve development interventions in two important ways. First, the process itself helps improve the quality of planning by assessing the quality of the needs analysis that has been undertaken, the plausibility of the planned intervention in terms of the needs and the stated goals, and the compliance of planning processes. Second, ex ante impact evaluation can contribute to better investment decisions by providing estimates of the benefits, cost-benefit ratios, and return on investment for potential projects through a model linking inputs with estimated outcomes and impacts.

The approach has been criticized for lacking a critical analysis of the intended benefits and their likely distribution (in particular whether the very poor are likely to benefit), for producing misleadingly precise results that conceal the assumptions and judgments underpinning the calculations, and for favoring larger projects over smaller ones (which might arguably be more locally appropriate) because of the high technical costs of conducting cost-benefit analyses in the ways recommended (Stewart 1978; Chambers 1978). It also does
not take account of the broader impact of development assistance which frees up government expenditure for other purposes (World Bank 1998).

Despite these concerns, ex ante evaluation continues to be widely used to inform investment decisions. Different terms are used for this approach, including ex ante impact assessment, return on investment (ROI), and value for money (VfM). A primary focus is on ensuring that investments are guided towards the projects most likely to produce benefits—and on reassuring donors that this is the case. These can vary from quick estimates of likely benefits, as in the poverty impact assessment approach developed by the OECD’s Development Assistance Committee (DAC) (1997), or more precise estimates based on econometric models. Another use for these models is to support simulations that can suggest changes to improve the reach or impact of projects, or reduce their cost without reducing their impact—for example, the ex ante evaluation of conditional cash transfer projects (Bourguignon, Ferreira, and Leite 2002). As development moved from discrete projects to more integrated and complex programs and broad policy reforms, and as aid moved from projects to more general budget support, the feasibility and relevance of this emphasis on ex ante project evaluation was further questioned.
Improving planning and accountability

Another early focus of development evaluation was to improve planning and management, especially by donor agencies. Much of this focused on the use of the logical framework, commonly referred to as the logframe. This is one of the most common and persistent approaches to evaluation in development, whose roots go back to the introduction of program budgeting into the U.S. Department of Defense in 1961.

Building on this experience, in 1965 a new Planning-Programming-Budgeting System (PPBS) was introduced across the U.S. federal government (Lyden and Miller 1967). In 1969, the U.S. Agency for International Development commissioned the company Practical Concepts to develop a guide applying the approach to the planning and evaluation of international aid projects. The approach centered around the use of a matrix, the logical framework, which was further developed for use by UN agencies by Gesellschaft für Technische Zusammenarbeit (GTZ), the German international development agency.

The logframe consisted of four components, each contained in a separate row of the matrix: activities, outputs, purpose (the rationale for producing the outputs), and goal (a higher-level objective to which this program and others contributed). For each component, four aspects were articulated, each in its own column: a narrative description, objectively verifiable
indicators, means of verification, and assumptions and risks (factors outside the control of the program on which the success of achieving that component depended).

In addition to its use for planning and developing performance monitoring indicators, the logframe also provided a framework for external evaluation. These external evaluations tended to use the same criteria as research, emphasizing comprehensiveness and rigor, rather than producing timely and relevant information to inform decisions. This way of framing and describing programs continues to be a key feature of development evaluation. While logframes, and the logical framework approach more broadly, have helped provide a structure for both planning and evaluation, there are concerns about their use. These concerns relate to:

“‘logic-less frames’, where only an illusion of logic is provided; ‘jamming’ of too much into one diagram; ‘lack-frames’, which omit vital aspects of a project; and ‘lock-frames’, whereby programme learning and adaptation are blocked.”

(Gasper 2000: 21–2)

Some changes have been made to the format of logframes—for example, the UK Department for International Development (DFID) has changed some of the terminology being used, and the level of detail required. However, it is not clear that these innovations actually address the most potent criticisms of the approach.
The logframe-inspired approach to evaluation has also been reflected in the broader results-based management (RBM) emphasis in the UN system and more widely. The 2005 Paris Declaration on Aid Effectiveness committed partner countries and donors to manage for results, “managing and implementing aid in a way that focus[es] on the desired results and uses information to improve decision making,” and to “work together in a participatory approach to strengthen country capacities and demand for results-based management.” The Accra Agenda for Action in 2008 prioritized a results focus to improve aid effectiveness, and the Busan Partnership for Effective Development Cooperation in 2011 again supported the principle of focusing on results. Development evaluation training, such as the International Program for Development Evaluation Training, has focused on incorporating evaluation as an integral part of results-based management, with its textbook *The Road to Results* (Morra Imas and Rist 2009).

Studies of RBM in practice, however, have found that its performance often falls short of its promise. A survey of members of the UN Evaluation Group in 2006 found that indicators and evaluations were rarely used as practical management tools, but were used mainly for reporting purposes. Studies of the UNDP and UN Secretariat found similar problems (Bester 2012).
Developing appropriate indicators and using them appropriately for management are not easy tasks. They require adequate resourcing of time, technical expertise (in content and monitoring), and legitimacy to overcome a tendency to focus on procedural compliance in terms of inputs, processes, and outputs. Perrin (2002), in his report of an OECD meeting on challenges to results-focused management, summed it up succinctly: “If the data that are reported do not reflect actual performance, they are meaningless or worse. It would be misleading or even dangerous to use them for policy- or decision-making.” Gaming, data corruption, and goal displacement are predictable risks that need to be directly addressed.

More recently, the need for credible evaluation of performance has increased with the introduction of “performance-based aid,” where governments or non-government organizations receive additional aid if they meet performance targets (Eichler and Levine 2009; Olken, Onishi, and Wong 2012).

**Rapid appraisal**

Traditional external evaluations, with their emphasis on rigor and comprehensiveness, often lacked focus and timeliness, which limited their usefulness. Rapid rural appraisal emerged in response to these concerns. While many of the shifts in approaches to development evaluation relate to debates and dynamics in the broader evaluation field, rapid rural appraisal
arose from development evaluation experience. Rapid rural appraisal grew out of Farming Systems Research and Extension experiences and techniques promoted by the Consultative Group on International Agricultural Research Centers (CGIAR). Agriculture development specialist Anthony Ellman developed rapid rural appraisal after undertaking a study of the achievements of four different land settlement schemes in Sri Lanka (Ellman 1981). It took nine months to collect data and six months to complete the 305-page report for the study, by which time it was no longer relevant, as government policy had changed. Ellman’s next evaluation took a very different form: after the minimum data required had been identified, a concise 25-page report was produced, on the basis of six weeks of data collection by a team of ten people, with clear recommendations that were largely accepted (Crawford 1997).

Rapid rural appraisal is not a single method of data collection, but an overall approach to focusing and prioritizing data collection, and then mobilizing a team to collect and analyze data quickly to produce a succinct and targeted report. It should not be confused with the “quick and dirty” approach sometimes used in development evaluation, where one or two evaluators produce a report on the basis of a review of existing documentation and a brief field visit, limited to interviews with readily accessible sites and informants. Robert Chambers (1980) labeled this superficial approach “rural development tourism” and criticized its inherent biases in data collection and analysis. Rapid rural appraisal, by contrast,
was intended to be “fairly quick and fairly clean,” and to use a larger team with the necessary expertise to ensure an adequate standard of data collection and analysis in a short time.

Unlike the other approaches discussed in this chapter, rapid rural appraisal has not continued to be used widely. Some of its elements were incorporated in participatory approaches, but it seems to have been largely ignored in more recent discussions of development evaluation, again reflecting the lack of learning from previous practice and the continuing emphasis on methodologies advocated by academics publishing in refereed journals.

**Participatory approaches**

While rapid rural appraisal addressed concerns about the timeliness of evaluations, it was still primarily focused on how external evaluators could provide information to inform decisions to be taken by external funders or policy-makers. With the increasing emphasis on the role of communities in producing sustained and effective development came a matching emphasis on involving communities in evaluation (Estrella et al. 2000) and a realization that their involvement in data collection could achieve greater coverage, more valid interpretation of data, and greater local ownership.
This trend reflected discussions in the wider evaluation literature about the value of involving project stakeholders in some way in an evaluation. Utilization-focused evaluation (Patton 1978, 2008) was based on two principles: identifying the primary intended users of an evaluation and their intended uses of it, and then making every subsequent decision about the design and conduct of the evaluation to ensure it was appropriate for their intended use.

Action research focused on supporting those implementing a service to engage in cycles of improvement through an iterative process of observe-reflect-plan-act (Kemmis and McTaggart 2005; Burns, Harvey, and Aragón 2012).

Some approaches to participation focused on engaging local project staff; others focused on community members or intended beneficiaries. This participation could be at one of three levels, depending on whether the underlying rationale for using a participatory approach was primarily to improve the quality and relevance of data collected or to engage participants in implementing changes.

The first level of participation involves participants (especially community members) as important sources of information about the implementation of projects and programs, their outcomes and impacts, and how well these match community needs. One approach to this level of participation, beneficiary assessment, was developed by the World Bank as “a process of listening systematically to key actors” (Salmen 2002) in order to document the
experiences and values of the people who were intended to benefit from programs and projects. The rationale for beneficiary assessment was that “by encouraging people to express their beliefs and values, [it] leads to development which responds to, while it promotes, the fuller participation of people in their own development” (Salmen 2002). Its primary intended purpose, however, was to inform decisions by managers responsible for the particular project or program.

Using a combination of direct observation, conversational interviews, and participant observation, beneficiary assessment developed a picture of the concerns and experiences of intended beneficiaries. This could be used to check if the goals of the project matched the aspirations and needs of the community, to understand barriers to engagement and implementation, and to gather evidence about the causal paths by which the project or program produced its results. By 2002, over three hundred projects funded by the World Bank in more than sixty countries had used the approach. However, in part because of the time needed, its widespread use has not continued, although most development evaluations seek to have some processes for including the “voice” of intended beneficiaries.

The second level of participation involves the community in collecting data. Participatory rural appraisal comprised a range of methods that could be readily used with a range of participants (including non-literate community members), such as community sketch maps,
group-generated timelines, and trend lines. Participatory data collection can improve the quality of data collected (where community members can gain better access to key informants) and improve the quality of data analysis (by appropriately interpreting responses). It can also build community engagement in a project (Chambers 1994). It is important not to conflate participatory data collection with qualitative methods. Many of the approaches used in participatory rural appraisal involve quantifying data, with a particular emphasis on ensuring the right things are counted in the right way in order to represent the situation accurately (Chambers 2007).

The third level of participation involves control over the evaluation process itself. Sometimes this involves participatory data collection as well, and other times it involves community members being part of the decision making process, overseeing the activities of external evaluators who collect and analyze the data.

Many further developments of participatory approaches have emphasized strengths-based approaches to evaluation to support strengths based approaches to development. The “positive deviance” approach, initially proposed by Marian Zeitlin and operationalized and developed by Jerry and Monique Sternin in the 1990s (Pascale, Sternin, and Sternin 2010), was developed for situations where, although the average experience is poor, there are some success cases that can be learned from. Rather than having an external evaluator conduct an
evaluation, identify lessons to be learned, and then try to convince community members to adopt them, the “positive deviance” approach involves the community in the entire process, including deciding what constitutes success. In an early example in Vietnam, researchers worked with community members to identify “outlier” families whose children were unusually well nourished and healthy, in a poor community where child malnutrition was widespread and seemingly intractable. Discussions with these families revealed differences in their feeding practices (adding crustaceans to the rice, providing several small meals) that could be adopted by other families. “Appreciative inquiry” takes a similar strengths-based approach, working with implementers to identify examples of good practices and explore ways of increasing their frequency (Elliott 1999).

Concerns have been expressed about the rigor and credibility of participatory approaches, especially when they have a strengths-based focus. In particular there are questions about possible bias in data collection because of participants’ prior involvement (Leeuw and Vaessen 2009) and concerns that such evaluations can too easily be co-opted to avoid hard questions. Skills in group facilitation are needed for effective use of these approaches.
Accountability to communities

Earlier approaches to accountability emphasized the upward accountability of recipient governments to donors, and of program managers to senior management and elected officials in government. To complement this there developed an interest in downward accountability to the communities. Most governments have some level of accountability to their communities in the form of elections, but even when elections are free and fair and voting provides a level of accountability to the government, this is a “long path” of accountability, with infrequent opportunities to provide overall feedback on the government’s performance.

Various evaluation approaches have been developed to provide a shorter path for accountability, each with its own logic of what constitutes good development and how evaluation can help to achieve it. Some approaches have focused on identifying gaps and shortfalls in service delivery, responding to basic problems such as teacher absenteeism, service opening hours, and availability of supplies. Citizen Report Cards, developed in the Indian state of Bangalore by the Public Affairs Centre, a local NGO, gathered user feedback on the performance of public services, such as whether health clinics were open at the advertised times. The process involved extensive media coverage and civil society advocacy to achieve change by “naming and shaming” poor performers (Ravindra 2004).
Technological developments have made it easier for communities to report on observable aspects of service delivery performance. The increasing use of mobile phones, especially in Africa, has provided a way to get information quickly from a dispersed team of data collectors or from community members. For example, the Stop Stock Outs project in Kenya conducted coordinated surveys of the drugs available at health clinics, reporting on those that did not have first-line medicines for malaria, HIV/AIDS, and other common diseases.

All of these projects depend on the willingness and ability of government agencies to respond to identified problems. Some approaches to community accountability work to help develop this capacity. For example, “Citizen Voice and Action,” developed by the international NGO World Vision, also includes processes to support dialogue between communities and service providers so they can together develop solutions to identified problems and monitor progress in addressing them, rather than taking an adversarial approach (Walker 2009).

**Finding out “what works”**

The increasing focus on improvement of programs led to concerns that hard questions were not being asked about terminating ineffective programs, and that insufficient attention was being paid to identifying successful technologies that lend themselves to being scaled up.
There was also concern that public support for international aid was faltering, requiring a demonstration of the value of investments in development. These concerns, combined with the rise of the “what works” movement in the USA and UK, led to systematic efforts to increase the use of experimental research designs in development evaluation. These efforts were explicitly inspired by the use of randomized control trials (RCTs) in clinical trials to test the effectiveness of different drugs or medical treatments, and they borrowed the concept of the “gold standard” and the “hierarchy of evidence,” where RCTs were seen as the highest quality research design, some quasi-experimental designs (propensity scores, difference-in-difference) as the second highest quality, and non-experimental designs as poor quality.

While RCTs had been used in social program evaluations since the 1960s, their use in development began much later. In 2003 the Poverty Action Lab at the Massachusetts Institute of Technology (MIT) and Innovations for Poverty Action based at Yale University were established with the aim of creating a network to both undertake and advocate for the use of RCTs. In 2004, the Center for Global Development in Washington convened an “evaluation gap” working group to examine why what it termed “rigorous impact evaluations” of social development programs were rare, and what could be done to change this situation.

The resultant working paper, “When Will We Ever Learn?” (Savedoff, Levine, and Birdsall 2006), identified the fact that much of development evaluation did not adequately address
questions of impact. Far too often it consisted of a “quick and dirty” evaluation planned and conducted towards the end of a project, drawing on project documents and monitoring data, interviews with key informants, and sometimes some cursory observation in a field visit and group interviews with beneficiaries, all of whom were well aware that the continuation of the project depended on its being seen to be effective. The paper argued that weak evidence of effectiveness was produced by these evaluations, given the absence of good quality measures of impacts and information about any negative impacts, little attention to the match between intended beneficiaries and those who had actually benefited, similar inattention to the likely sustainability of these impacts, and little effort at attribution beyond noting that the intended impacts had occurred. These discussions led in 2009 to the launch of the International Initiative for Impact Evaluation (3IE), an organization created to generate, synthesize, and support the use of rigorous evidence to inform policy in development.

Advocates for the approach claimed it was particularly well placed to be able to identify interventions that work well so that they can be scaled up and rolled out more widely (Duflo 2004; Duflo, Glennerster, and Kremer 2007). Its particular strength was construction of a counterfactual: a group that does not receive the intervention that is to be evaluated, but that can be compared statistically to the group that did receive the intervention. A central plank in the case made by proponents of randomized control trials is that by randomly assigning subjects to either to the experimental or the control group (i.e., the group that
receives the intervention or the group that does not), the only difference between them will be
the intervention, and thus differences between the two groups can logically be ascribed to the
intervention and the intervention alone. Yet RCTs need to manage the risk that random error
will produce non-equivalent group needs. This can be achieved through large sample sizes,
inspecting groups for comparability on relevant variables, and conducting multiple studies.

The RCT approach to evaluation has been criticized on a number of grounds. Concerns have
often been expressed about the ethical implications of experimental research designs, where
only some participants receive a potentially beneficial program. This can however be readily
addressed, especially in resource-constrained situations, by ensuring that the control group is
queued to receive the program if it is shown to be effective. Other critiques have highlighted
the fact that evidence-based medicine does not use RCTs alone, but draws on both quasi-
experimental research (epidemiological designs) and non-experimental designs (including
single case studies). RCTs have also been criticized for their “misunderstanding of
exogeneity and the handling of heterogeneity” (Deaton 2010).

Some concerns have related to the methodological aspects of the approach and its ability to
produce useful knowledge, given its assumption that an estimate of mean differences in
outcomes is the most useful finding for informing policy and practice (Heckman 1992). A
key concern has been that the emphasis on the average effect risks masking differential
effects, which can be large and important. A narrow focus on “what works” does not distinguish between the winners and losers from a particular project or policy, nor identify whether the result is desirable.

For example, a microfinance project might have a slightly positive impact, on average: by producing strong positive impacts for less poor people who are better able to put the money to productive use to generate income, and strong negative impacts for very poor people who are more likely to access money to meet immediate needs. Very poor people lack the capacity to use loans to generate income to improve their situation or even repay the principal. While it is sometimes possible to identify subgroups whose experience is likely to be different from the average—men and women, urban and rural, indigenous and other people—, in other cases these subgroups are not evident in advance, being based on factors other than simple demographics (for example, the level of family violence in the household). Because the statistical power of RCTs is designed with these identified subgroups in mind, it is not always possible to disaggregate them by factors that become evident only during the implementation of the project.

Some concerns relate to the practical implementation aspects of RCTs and their claims of rigor. Achieving the high standards required can be challenging. In many instances, comparison groups are used instead of control groups, or the process of randomly assigning
participants to different groups is poorly done, because of high attrition rates or other breakdowns in the experimental design. Whatever the reason, if random assignment is poorly done, its key purpose, which is to address selection bias, is not achieved and the ability to demonstrate causal attribution is drastically reduced. Deaton (2010) has concluded that “actual experiments are frequently subject to practical problems that undermine any claims to statistical or epistemic superiority.”

Another concern has been that the focus on developing evidence of what works and convincing donors and governments that they should support certain initiatives is not helpful when the focus of agencies is to improve programs to which a formal (or political) commitment has already been made. This has highlighted the need to work on the “demand” end of evidence-based policy as well as on the “supply” end. Demand includes the individual skills and motivations of policy makers and other decision makers to request evidence to support their decisions, and the political and organizational environment that either supports or inhibits this—for example, requirements for evidence to be systematically included in planning proposals. Supply includes the skills and motivations of evaluators, researchers, and community members to generate appropriate evidence to inform decisions.

Concerns have also been expressed about the limited range of interventions suitable for experimental research designs. RCTs are at their best in discrete projects with a large
population size, where randomization is an option and where impacts will be discernible within a reasonably short time. Development interventions that operate at the level of the community, region, or nation do not lend themselves to using random assignment, as the unit of analysis is often too large (thus masking internal variation) and too few in number (thus negating meaningful comparisons).

More recently, increasing attention has been paid to the limitations of single studies, however large and well-designed and -implemented. Results from a single study can still be subject to random error, which random assignment does not reduce, and its external validity, or its ability to generalize findings from one site or population to another, is limited. As a result, there is increasing attention on conducting multiple studies and undertaking systematic reviews that summarize the range of credible evidence on a topic. This has included creating a new international development coordinating group in the Campbell Collaboration, an organization that focuses on supporting systematic reviews of evidence in social programs.

Some systematic reviews include only evidence from RCTs, despite concerns raised in evidence-based medicine about the risks of excluding potentially relevant evidence from other research designs.¹ Other reviews include some quasi-experimental designs, such as propensity score matching and regression discontinuity. Evidence can be synthesized using meta-analysis, producing an overall statistical summary, or narrative synthesis along the lines
of a literature review. Meta-analyses cannot successfully address contextual variations adequately unless all the studies have presented disaggregation of their data using the relevant variables.

Another significant development in this approach has been increasing interest in using causal inference techniques from natural science and social science, such as multiple lines and levels of evidence, process tracing, comparative case studies, qualitative comparative analysis, and contribution analysis (Iverson 2003; Stern et al. 2012) to identify “what works” in situations where experimental or quasi-experimental designs are not possible. More radically, some approaches argue that, given the multiple contributors to impacts, it is not possible to undertake credible attribution of impacts to an intervention. Outcome mapping (Earl, Carden, and Smutylo 2001) focuses on identifying “boundary partners”—organizations outside the control of an intervention but within its sphere of influence, and critical to achieving the intended impacts—and articulating progress markers of intermediate outcomes related to their activity.
What works for whom

Realist evaluation (Pawson and Tilley 1997) has been increasingly used in health and community services evaluations, particularly in the UK and Canada. Its use in evaluation addresses some of the limitations of the “what works” agenda. A realist evaluation asks: What works for whom, under what circumstances, and how? This goes beyond simply identifying and reporting patterns of results. Realist evaluation involves iteratively constructing and testing (against a variety of evidence) “context-mechanism-outcome configurations”—different theories about causal mechanisms that operate in particular contexts (implementation environment or participant characteristics) to produce certain outcomes. The strength of this approach is its investigation of heterogeneity and external validity, producing results with more information about their potential for transferability to other sites. However, the approach has been criticized for providing less clear messages for policy makers.

A further development of this approach has been realist synthesis (Pawson 2006), a systematic review using a realist framework. It draws on a range of credible evidence, including impact evaluations that use non-experimental research designs, and seeks to explain how the evidence reflects causal mechanisms that work differently in different contexts. These reports lend themselves to moving beyond knowledge collection to knowledge
translation—helping policy makers and practitioners understand the situations in which interventions are likely to be effective, and the types of adaptation that will be appropriate.

**Responding to complexity**

At the same time as the rise of the RCT-oriented understandings of evidence-based policy, other researchers seeking to inform policy were exploring how insights from complexity science could be used to help individuals, communities, and organizations develop the resilience needed to respond to local circumstances and ongoing change (Ramalingam et al. 2008). Interestingly, given the emphasis on drug trials as an exemplar for research and evaluation, many of the examples of applying complexity concepts in evaluation have been undertaken in the health area (Sibthorpe, Glasgow, and Longstaff 2004).

Complexity-based approaches recognize the rapid, continuing, and discontinuous nature of change, which can be poorly served by evaluation approaches that are premised on a fairly stable situation where what works largely remains constant and evaluation is mostly about checking compliance with plans. In the complexity-based approaches, effective development is seen not as identifying what works and then scaling it up, but rather as supporting ongoing
adaptation and responsiveness to new needs, risks, and opportunities, and to individuals’ different needs.

Complex-adaptive approaches to development evaluation draw on Glouberman and Zimmerman’s (2002) distinction between simple, complicated, and complex. In this taxonomy, “simple” describes situations in which there are rules that can be followed and there is agreement on a single way to achieve a result; “complicated” indicates the presence of many components, each of which requires expertise and coordination; and “complex” indicates the presence of emergent and responsive features, with relationships being the key to success. Kurtz and Snowden (2003) characterized simple as the domain of the “known,” where cause and effect are well understood, and best practices can be confidently recommended. They see complicated as the domain of the “knowable,” where expert knowledge is required, and complex as the domain of the “unknowable,” where patterns are evident only in retrospect. Simple interventions need competent transfer of findings. Complicated interventions need communication of differentiated findings, and support to select and adapt them appropriately. Complex interventions need support for ongoing learning.

Dynamic and emergent interventions present a challenge to conventional linear processes of developing an evaluation, implementing it, and reporting the findings. Dynamic interventions
will change substantially over time, and their specific impacts cannot always be identified in advance. For example, community development projects often work by bringing together community members to identify local issues, develop a plan for addressing them, gather resources and implement the plan, and then review progress and plan again. Where projects have a similar desired outcome, they are unlikely to implement the same activities.

Developmental evaluation (Patton 2010) aims to support ongoing adaptation and learning to respond to emerging opportunities and priorities and to the different needs of individual communities or beneficiaries.

Because complexity-based approaches to evaluation are relatively new, there are still few examples of the approach or analyses of its strengths and weaknesses. There has been concern that evidence generated by such evaluations is insufficiently clear and does not provide either policy-makers or program implementers with clear direction or instruction on what should constitute good practice. A related concern appears to be that this approach can provide an overly persuasive reason not to do good, rigorous planning, and that generalized rough frameworks may suffice. There is also a concern that it assumes a high level of skills and knowledge among service deliverers and a commitment to ongoing learning, assumptions that might not be appropriate in many resource-limited situations.
Country-led evaluation

The Paris Declaration on Aid Effectiveness (OECD 2005) outlined “far-reaching monitorable actions” to reform how aid is delivered and managed. The overarching concern was to strengthen governance and improve development performance. The declaration directly addressed the shift in how development should be curated. Previously, the underlying assumption was that much of the work being done in development was focused on meeting the expectations of donors and providing evidence to be accountable to the North. The Paris Declaration shifted this explicitly and made it clear that developing country processes should belong to and be owned by the country concerned. The corollary is that evaluation should also be for use by the government responsible for the country.

More recently this trend has been accelerated by the strong growth in the developmental state, or state capitalism. In this model, the state makes highly deliberate and purposeful use of market mechanisms to achieve developmental outcomes. The prime examples of this approach are China and Brazil. In a developmental state, the focus shifts from discrete, individual evaluations that provide evidence to external donors for accountability purposes, to a greater concern with evaluation systems that generate useful knowledge for use in informing policy and improving program implementation. The long-term implications of this
shift have yet to show themselves. Associated with this is increasing emphasis on building national evaluation capacities, in terms of supply and demand, including the skills and expertise of internal and external evaluators and evaluation users, as well as organizational incentive systems. Evaluation associations can play a role in both of these (Holvoet, Dewachter, and Gildemyn 2011).

**Conclusion**

The different types of evaluation are summarized in Table 1, in terms of their underlying assumptions about what constitutes development, and how evaluation can support this. Effective development evaluation matches the approach to development that is being used, and how evaluation is intended to support it. The various approaches to development evaluation discussed in this chapter differ in important ways in terms of:

- What should be the main focus of evaluation: pilot projects or all projects; discrete projects or broader policy initiatives;

- When evaluation is most important: before, during, or after implementation;
• Who should be involved in doing evaluation: internal staff, external evaluators, community members;

• Who should control it: individual donors, a joint group involving partner governments, or an alliance of donors, national governments, and community groups;

• What relative emphasis should be placed on monitoring versus evaluation;

• Whether the methods and standards for evaluation should be the same as to those for research.

An awareness of the different approaches, and the types of development for which they are best suited, increases the likelihood that evaluation will be appropriately matched to its situation and succeed in contributing to development.
Table 9.1: Different types of evaluation

<table>
<thead>
<tr>
<th>If development is mostly about</th>
<th>And evaluation supports development by</th>
<th>Then the type of evaluation needed will be</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>PROVIDING INFORMATION FOR DECISION-MAKERS</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Choosing the right programs to invest in</td>
<td>Helping to manage investment risk</td>
<td>Ex ante impact evaluation</td>
</tr>
<tr>
<td>Effective planning and management of projects and programs</td>
<td>Helping to clarify what needs to be done and providing ongoing feedback on progress</td>
<td>Performance monitoring (logical framework analysis, results-based management); external review; rapid rural appraisal</td>
</tr>
<tr>
<td>Scaling up effective projects and programs</td>
<td>Identifying “what works” and monitoring fidelity of implementation</td>
<td>Experimental or quasi-experimental evaluations that provide estimates of average net impact; compliance monitoring of activities</td>
</tr>
<tr>
<td>Translation and adaptation of appropriate technology</td>
<td>Identifying what works in what circumstances and supporting implementers in translating findings to new situations</td>
<td>Explanations of how interventions work, and under what circumstances (“what works when,” “good practices,” realistic evaluation)</td>
</tr>
<tr>
<td>Resilience in the face of uncertainty and rapid change</td>
<td>Supporting ongoing adaptation and responsiveness</td>
<td>Real-time evaluation to support front-line workers; dialogue between partners</td>
</tr>
<tr>
<td>Supporting local people in becoming agents of their own development</td>
<td>Supporting beneficiaries and other stakeholders in developing and sharing solutions, including managing, conducting, and using evaluation; learning from success</td>
<td>Participatory approaches; strengths-based approaches; building national capacity for evaluation</td>
</tr>
</tbody>
</table>
CHANGING BEHAVIOR THROUGH INCENTIVES

<table>
<thead>
<tr>
<th>Donors and central government ensuring that national partners (government agencies and NGOs) do the right thing</th>
<th>Identifying those not doing the right thing so they can be sanctioned, and motivating all to do the right thing</th>
<th>Upwards accountability</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Identifying those doing the right thing so they can be rewarded, and motivating all to do the right thing</td>
<td>Rewards systems (performance-based aid)</td>
</tr>
<tr>
<td>Civil society ensuring that government agencies and NGOs do the right thing</td>
<td>Identifying and sanctioning those not doing the right thing and/or supporting agencies to improve their performance</td>
<td>Community accountability</td>
</tr>
</tbody>
</table>

*Source: author’s own design (Patricia Rogers and Dugan Fraser).*
References


---

1 This point was made eloquently by Smith and Pell (2003) in their review of evidence for the effectiveness of parachutes as a technology to reduce trauma from jumping out of planes.