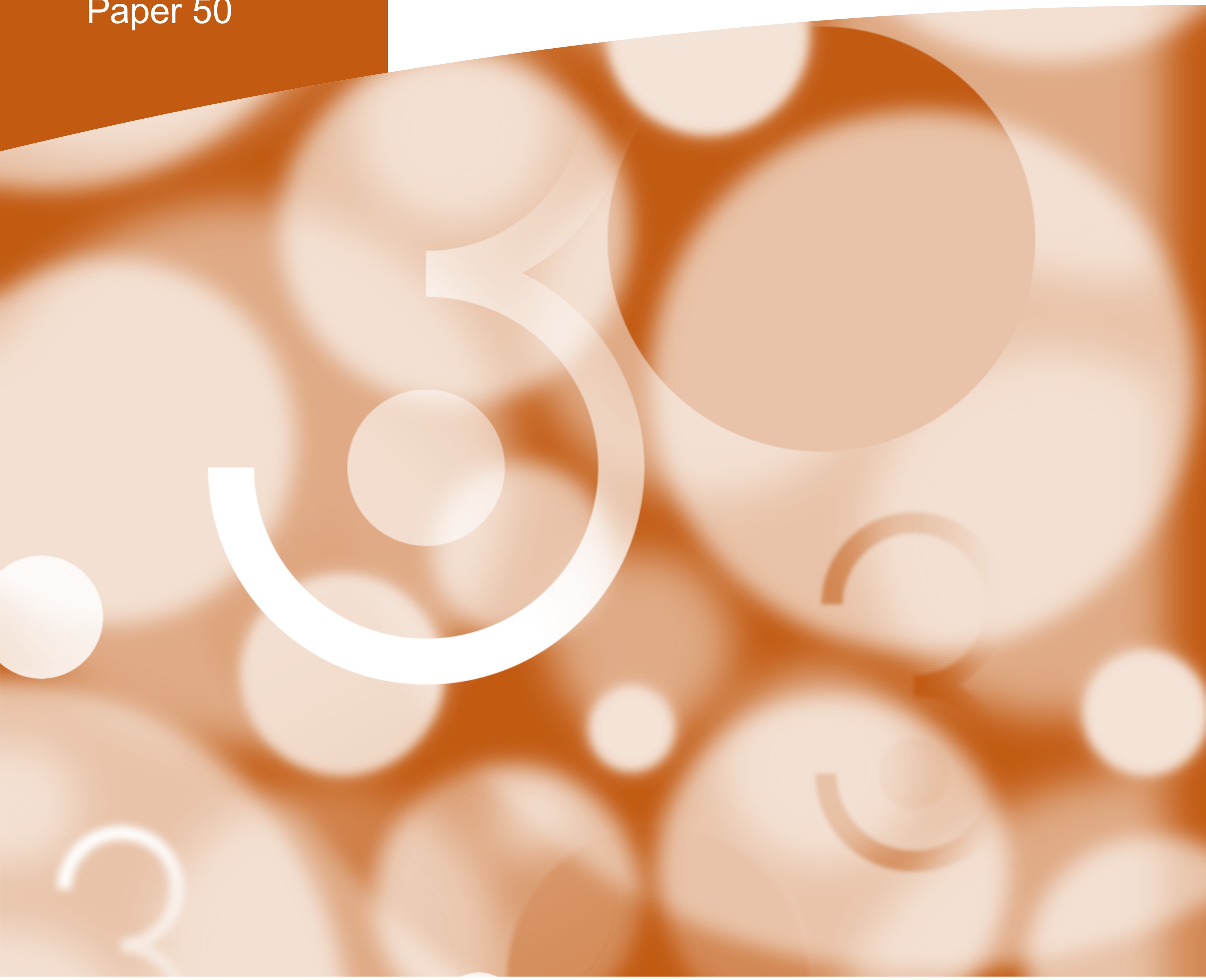
Vibecke Dixon
Michael Bamberger

Incorporating process evaluation into impact evaluation

What, why and how

March 2022

Working
Paper 50



International
Initiative for
Impact Evaluation

About 3ie

The International Initiative for Impact Evaluation (3ie) promotes evidence-informed equitable, inclusive and sustainable development. We support the generation and effective use of high-quality evidence to inform decision-making and improve the lives of people living in poverty in low- and middle-income countries. We provide guidance and support to produce, synthesise and quality assure evidence of what works, for whom, how, why and at what cost.

3ie working papers

These papers cover a range of content. They may focus on current issues, debates and enduring challenges facing development policymakers, programme managers, practitioners and the impact evaluation and systematic review communities. Policy-relevant papers in this series synthesise or draw on relevant findings from mixed-method impact evaluations, systematic reviews funded by 3ie, as well as other rigorous evidence to offer new analyses, findings, insights and recommendations. Papers focusing on methods and technical guides also draw on similar sources to help advance understanding, design and use of rigorous and appropriate evaluations and reviews. 3ie also uses this series to publish lessons learned from 3ie grant-making.

About this working paper

This paper, *Incorporating process evaluations into impact evaluations: what, why and how*, consists of guidelines that can provide impact evaluators with tools and ideas on how to explore and add relevant elements of process evaluations to experimental and quasi-experimental impact evaluation designs.

The content of this paper is the sole responsibility of the authors and does not represent the opinions of 3ie, its donors or its board of commissioners. Any errors and omissions are also the sole responsibility of the authors. All affiliations of the authors listed on the title page are those that were in effect at the time the paper was accepted. This document has been copyedited as formatted as per 3ie style.

Suggested citation: Dixon, V, Bamberger, M, 2021. *Incorporating process evaluation into impact evaluation: what, why and how,* Working Paper 50, New Delhi: International Initiative for Impact Evaluation (3ie). Available at: DOI http://doi.org/10.23846/WP0050

Incorporating process evaluation into impact evaluation: what, why and how

Vibecke Dixon
International Initiative for Impact Evaluation (3ie)

Michael Bamberger
3ie

Working Paper 50

March 2022

About these guidelines

These guidelines focus on how process evaluations may strengthen the outcomes of impact evaluations[1] and, in turn, improve the design and implementation of ongoing and future development interventions.

The present guidelines have been developed to provide impact evaluators with tools and ideas on how to explore and add relevant elements of process evaluations to experimental and quasi-experimental impact evaluation designs. The guidelines can be applied to both (1) retrospective impact evaluations conducted at the time of project completion or even one or two years later and (2) prospective impact evaluations that combine pretest and posttest comparison designs.

The guidelines point to common pitfalls in project design and implementation that, if not identified at a sufficiently early stage, could result in misleading conclusions in impact evaluations and systematic reviews of what works and not in efforts to advance development. Many impact evaluations implicitly assume that projects were appropriately designed and implemented as designed, thus frequently ignoring potential misdiagnosis in the project design or lack of fidelity—i.e. significant deviations from the implementation protocol. It is our contention that smartly designed and rigorous process evaluations can add valuable new knowledge that facilitates sharper tailoring of impact evaluation designs to the real-world context within which projects are implemented, and their ability to detect the *whys* and *hows* of the intervention results.

It is recognized that different project designs and implementation arrangements, as well as a variety of contexts may produce disparate scenarios that require flexibility in evaluation design and execution, depending on distinct information needs. The guidelines emphasize that understanding project design and implementation requires the combination of different sources of information to measure a range of implementation factors, and monitor the complex socio-cultural and political factors that influence implementation processes. While most impact evaluators already recognize this, the guidelines recommend that the process of combining different data sources should be made more explicit by the use of a mixed methods evaluation framework. This framework also strengthens evaluation rigor through a systematic use of triangulation, where validity is strengthened by comparing three or more independent estimates of key indicators. These guidelines, therefore, outline a variety of mixed-method designs that may be applied depending on information need and context. They are meant to help evaluators ensure that important and relevant elements receive appropriate attention in evaluations, not to provide a “recipe book” for how to (always) do it.

The process evaluation framework developed for these guidelines comprises four main elements:

1. Design
2. Implementation
3. Institutional aspects
4. External and contextual factors, which are underlying the three previous elements

[1] Here, impact evaluations are defined as rigorous studies that measure the effects of international development programs. They measure changes in development outcomes that can be attributed to a specific development intervention through a credible counterfactual.

The main messages in these guidelines are:

- **Process evaluation can help strengthen impact evaluation designs** by explaining the intervention's positive, negative, significant, insignificant or unexpected results. Understanding the mechanisms for how and why development interventions produce successful change, or fail to do so, is key to refining theory of change and improving intervention effectiveness. A key purpose of process evaluations is to carefully assess the different factors that could influence project outcomes.
- **Implementation assessment** is already used in many impact evaluations, and the purpose of process evaluation is to provide time and resources to permit the use of a wider range of data collection and analysis tools and techniques.
- **Attention to possible misdiagnosis and prescription of the wrong treatment is crucial to avoid misleading conclusions about what works and what does not.** Assessment of the intervention's problem analysis (which is underlying the project design) is key. So far, this has not been commonly included in existing process evaluation guidelines, in spite of the serious consequences for project outcomes of possible misdiagnosis and failure to identify the core problem at the project design stage. A weak problem analysis might lead to misleading interpretations and explanations of reasons a project succeeded or failed to achieve its objectives; this, in turn, may result in misguided conclusions in both impact evaluations and systematic reviews.
- **Attention to both the factual and the counterfactual can provide valuable new insights into strengths and weaknesses in the implementation process.** Many experimental and quasi-experimental impact evaluations focus on design and finding a credible counterfactual and implicitly assume that the project was implemented as planned. If the evaluation finds no statistically significant differences between the project treatment and comparison group, it is assumed this was due to a weak theory of change or project design failure. It is equally plausible that the cause is misdiagnosis (failure to identify core problem) or weak project implementation.
- **Process analysis can detect groups that are excluded, underserved or in some cases, worse off as a result of the project intervention.** These mechanisms of exclusion are often difficult to detect and may require more in-depth qualitative approaches than normally possible within an impact evaluation.
- **The evaluation design should match the information need**, not the other way around. Evaluations should be issues driven, not methods driven. It is necessary to first identify the stakeholders' information needs, then select evaluation methods based on what kind of information each method can provide. This will often require a mixed-methods approach.
- **Application of mixed-methods approaches,** where qualitative and quantitative methods are combined to permit the evaluator to draw on the widest possible range of evaluation methods and tools, increases the validity of conclusions by providing three or more independent estimates of key

indicators (triangulation). It also permits a deeper and richer analysis of interpretation of the context a project operates in, and offers ways to reduce the costs of time and data collection.

5. **Triangulation is a key element of mixed methods.** This requires that three or more independent estimates be obtained for all key indicators, and that there be a mechanism to compare the estimates to validate the approximated values.

Contents

List of figures and tables

Acronyms

3ie	International Initiative for Impact Evaluation
DFID	UK Department for International Development
MCC	Millennium Challenge Corporation
NGO	Non-governmental organization
OECD/DAC	Organisation for Economic Co-operation and Development/Development Assistance Committee
PRA	Participatory rural appraisal
QUAL	Qualitative evaluation methods
QUANT	Quantitative evaluation methods
RCT	Randomized controlled trial
SHG	Self-help groups
UN	United Nations
UNICEF	United Nations Children’s Fund
WARDA India	Women’s Advancement in Rural Development and Agriculture Program,

1. Introduction

1.1 What is a process evaluation?

The definitions of process evaluation are many and varied, as outlined in Annex 1. The concept is used in a variety of ways and defined differently by diverse development organizations, evaluators and researchers.

For the purpose of strengthening impact evaluation designs, the core function of a process evaluation is that it may help explain the positive, negative, significant, insignificant or unexpected results of a development project or initiative. This is sometimes defined as helping to understand the who, what, where, when and how of project effects. It can also be used by managers to improve the performance of ongoing and future development interventions.

> The core functions of a process evaluation are to help explain positive, negative, significant, insignificant or unexpected results of an intervention, and to use this information to help managers improve the performance of the ongoing or future programs.

While an impact evaluation seeks to quantify project outcomes and assess the extent to which the observed outcomes can be attributed to the project intervention, process evaluation helps understand how the way in which the project was designed and actually implemented (compared to the proposed project design and implementation plan) may have affected the outcomes. Process evaluation can also help explain how the economic, political, socio-cultural, organizational and administrative factors that form the context where implementation takes place affect the outcomes. There is considerable conceptual overlap between process evaluation and formative evaluation, which many evaluation textbooks define as one of the two main kinds of project evaluation, the other being summative or impact evaluation.

Process evaluation is not a new concept. More than 50 years ago, Suchman's textbook on project evaluation provides a still valid and useful description of process evaluation:

> In the course of evaluating success or failure of a program, a great deal can be learned about how and why a program works or does not work. Strictly speaking, the analysis of the process whereby a program produces the results it does is not an inherent part of evaluative research. An evaluation study may limit its data collection and analysis simply to determining whether or not a program is successful (…). However, an analysis of process can have both administrative and scientific significance, particularly where the evaluation indicates that a program is not working as expected. Locating the cause of the failure may result in modifying the program so that it will work, instead of its being discarded as a complete failure (Suchman 1966, p. 66).

While Suchman emphasized the usefulness of process evaluations for identifying the causes of project failure, process evaluations are equally important when the expected outcomes are obtained. They help identify and understand the reasons behind a project's success that may provide crucial information for designers and implementers of current and future projects. However, it is possible that positive results materialized for other than the expected reasons. It is important to understand the reasons for successful outcomes, not just to explain reasons for failure. Importantly, process evaluation will often identify a number of additional outcomes (positive, negative or undefined) that were not included in the project design and are not integral to or captured in the impact evaluation.

Well-designed and well-executed process evaluations have the potential to significantly enhance the understanding of catalysts and barriers to implementation and other determinants of a project's success, thereby improving development interventions and practice. Understanding the mechanisms for how and why these interventions produce successful change, or fail to do so, is key to refining theory of change and improving intervention effectiveness. Process evaluation can help disentangle the effects of each element of the design and implementation and clarify the possible interactions that can occur to produce a synergetic effect. This is done by identifying and analyzing the various internal and external elements of a project design and implementation that may have an effect on project outcomes, as illustrated in *Figure 1*.

Figure 1: The various elements that may affect project outcomes

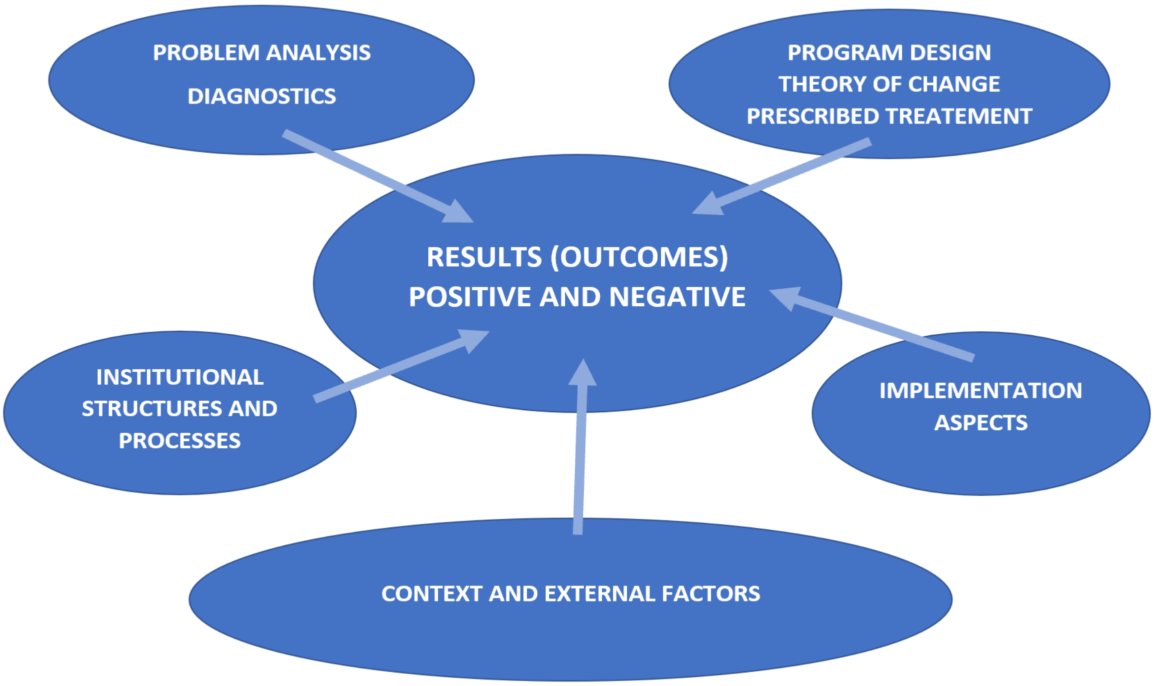


Figure 1

may cause positive or negative project outcomes. It is the process evaluators' task to identify and analyze these elements, and assess their effects on the outcomes. Without a process analysis, it is difficult to know whether the failure to achieve intended outcomes was due to i) project design failure, ii) project implementation failure, iii) institutional structures or iv) external factors beyond the control of both the designers and the implementing agency. Furthermore, many impact evaluations mainly rely on reports and data from the project agency. While this covers the participating project population, it often fails to identify sectors of the intended target population (e.g. the landless, illegal squatters, ethnic minorities) who may have been (intentionally or inadvertently) excluded, are not well served or may even be worse off as a result of the project.

Design failure, (i.e. the first category of failures identified above), occurs when the project design was inappropriate or inadequate to achieve the intended impact (i.e. there were flaws in the problem analysis, the theory of change or the design). In contrast, implementation failure (second category of failures identified above) happens when the project design was good, but there were problems with the way it was implemented. The third category refers to institutional structures—policies, financial arrangements, and institutional capacity and inter-agency collaboration, while the fourth category pertains to the multiple external (economic, political, socio-cultural, demographic or environmental) factors. Many of these factors are largely beyond the control of the implementing agency, but can be important in explaining why projects with identical designs can have significantly different outcomes in different locations. In practice, all four causes may be present to some degree. Each element and ways to identify and assess it will be presented in more detail in the following paragraphs.

Until now, assessments of the interventions' original problem analysis have not been commonly included in process evaluation guidelines, but we would argue that attention to possible misdiagnosis and prescription of the wrong treatment is crucial in any process evaluation, because any flaws in the problem analysis might have devastating effects on project outcomes. A weak problem analysis may lead to misleading interpretations of the reasons a project succeeded or failed to achieve its objectives. Real-world examples are presented in Chapter 2 *below*.

The present guidelines recommend that quantitative impact evaluations incorporate elements of process evaluation by applying a mixed-methods design that combines quantitative and qualitative research methods to capture breadth and generalizability (quantitative methods), as well as depth and rich description (qualitative methods). This combination provides abundant detail and multiple perspectives of project outcomes that neither quantitative nor qualitative methods can achieve on their own. In addition to the quantitative methods frequently applied to impact evaluations, participatory qualitative methods—e.g. observation and open-ended interviews and focus group discussions with a range of stakeholders—are often employed in process evaluations. Mixed methods, through a creative combination of multiple data collection methods and analysis, can also track ongoing, dynamic processes and provide a longitudinal perspective that covers a much longer time period than usually possible with most evaluation designs.[2]

There are three main process evaluation design options for incorporation into impact evaluations: (1) retrospective (ex-post), (2) pretest–posttest comparisons (baseline and end-of-project), and (3) dynamic or continuous observation over time. The latter can be linked to a theory of change where one clearly defined purpose of the process evaluation would be to test what Iversen and Lanthorn (2015) describe as critical assumptions in the project's theory of change.

[2] For example, some of the most insightful studies of the impacts of microcredit programs on women's economic and social empowerment have followed a small sample of households over very long loan cycles to observe the gradual process of change that unfolds over time, often years. While such a time horizon is not feasible in most evaluations, it provides a depth of understanding shorter term studies will miss. So, for organizations such as 3ie and other development agencies that have multi-year research programs in key sectors (such as women's empowerment), it is paramount to find creative ways to capture these longer time horizons, perhaps in collaboration with local non-governmental organizations (NGOs) or universities.

1.2 Elements of process evaluation are already familiar to impact evaluators

3ie and other development agencies already use in their impact evaluations many of the process evaluation tools and techniques described in these guidelines.[3] However, a number of these techniques are conducted under budget and time constraints, and the purpose of the guidelines is to provide a set of more intensive strategies that can be used when (1) adequate time and resources are budgeted and (2) importantly, when it is possible to be able to collect primary data over a longer period of time and in more depth than usually. This will require that process evaluation become a more integral part of an organization's impact evaluation strategic planning.

A full-scale process evaluation can be very resource intensive. Consequently, most organizations will need to apply the approaches selectively, either by only using an in-depth analysis in selected, priority evaluations or by focusing on less resource-intensive techniques. Part of this strategic approach can include pilot-testing the strategies in selected impact evaluations to assess the value added and get a clearer understanding when, where and how process evaluation can be justified.

1.3 How can process evaluations strengthen impact evaluations?

The previous section showed that 3ie and other organizations already incorporate many process evaluation techniques into their impact evaluations, but often with time and resource constraints. Therefore, the question addressed in these guidelines is:
What information can we get from a more intensive application of process evaluations and how would it be complementary to an impact evaluation?

As pointed out by Bamberger and colleagues (2006:205) and Bamberger and Mabry (2020), many quantitative evaluations use a pretest–posttest design, where the purpose is to assess the quantitative effects of the project (also called summative evaluation). With many of these designs, data are collected only at the start (pretest baseline data) and the end of the project (posttest); no information is collected during project implementation. The limitation of these designs is that they do not look into how well the project was implemented or how the implementation process affected project's outcomes or its accessibility to different sections of the target population. This lack of understanding the implementation process and context can be a particular problem in situations where it is necessary to know why a project did or did not achieve its intended outcomes.

[3] This introduction to 3ie's impact evaluations provides links to a number of 3ie impact evaluations that include some of the tools of process evaluation described in these guidelines: https://www.3ieimpact.org/What-we-offer/impact-evaluation. Some of the examples include: (1) The Indian National Rural Livelihoods program, (2) the Impact of Food Assistance on Food Insecure Populations During Projects in Mali, (3) Unpacking the determinants of entrepreneurship development and economic empowerment for women in Kenya and (4) Integrating impact evaluation and implementation research to accelerate evidence-informed action (with respect to HIV and AIDS). Some of the process evaluation techniques used include: (1) constructing ordinal scales for rating women's opinions of project impacts on the economic and social empowerment inside and outside the household), (2) using photos as a visual robustness check on whether small businesses appear to have grown over a three-year period, (4) focus groups and key informant interviews to inform the evaluation design, and (5) using 5-point scales to assess adherence to project requirements for approving loans.

Applying elements from process evaluation may provide different and additional information to an impact evaluation, such as detecting the underlying reasons an intervention may or may not have produced the intended outcomes. It could uncover the underlying assumptions or logical gaps in project design that caused a misdiagnosis or mistreatment, and it might detect any deviations in implementation from the intended design that may have direct effect on the intervention's results. As pointed out by Linnan and Steckler (2002), when interventions lead to significant outcomes, it is also important to understand which components of the intervention contributed to the success. A well-designed process evaluation could also help identify a broader range of positive and/or negative outcomes that were not addressed in the quantitative impact evaluation.

Process evaluations may serve as part of larger project evaluations. Their main purpose is to determine whether the project design was appropriate to address the identified problem, and whether and to what degree implementation was sufficiently close to design for an impact evaluation to be informative about the intended project. 3ie's guide on impact evaluation practice[4] states that "studies should clearly lay out how it is that the intervention (inputs) is expected to affect final outcomes, and test each link (assumption) from inputs to outcomes (sometimes referred to as the project theory). The evaluation design should incorporate analysis of the causal chain from inputs to impacts."

This is to avoid any incorrect conclusions in impact evaluations and systematic reviews about what does and does not work in development interventions, due to undetected misdiagnosis in design or factors during implementation. Also, when interventions lead to significant outcomes, it is important to understand which components of the interventions contributed to success,[5] so they may be scaled up and replicated. Thus, process evaluation assesses the quality and accuracy of the intervention as delivered to project participants or beneficiaries. Process evaluation can also make a critical contribution to discussing for whom different components of the project work and do not work, and how they work.

> The main purpose of process evaluation is to determine whether the program design was appropriate to address the identified problem, and whether and to what degree implementation was sufficiently close to design for an impact evaluation to be informative about the intended program, its results and the underlying reasons thereof. Process evaluation can also help understand the influence of the program's context on outcomes.

There are other benefits of process evaluation that are often not emphasized, but are important for strengthening social policies. The first such benefit is assessing the extent to which different outcomes can be attributed to the effects of the intervention. Many evaluations are used to validate a project's theory of change, and evaluators and managers may be too willing to accept results that support their theory of change. Even if a link can be established, it is important to know whether the intervention was a necessary and sufficient cause of the observed changes. Are there other conditions that must also be present and what are the conditions under which similar results could be

[4] Available at: https://www.3ieimpact.org/evidence-hub/publications/working-papers/theory-based-impact-evaluation-principles-and-practice

[5] Linnan and Steckler, 2002. Process Evaluation for Public Health Interventions and Research, p. 2.

expected in future projects? The second benefit is the important question of social exclusion. Often, the aggregate results of a project may be positive, but the evaluation may not have the necessary tools to identify any social, economic or ethnic groups that are excluded, either intentionally or inadvertently.

A process evaluation would, thus, identify:

1. To what extent the project effects (or lack thereof) were due to project design or implementation
2. How the institutional process and structures of project delivery might have affected project outcomes
3. How and to what extent external factors affected the results
4. Which parts of the project worked for whom
5. Whether any groups were excluded from project benefits or faced barriers that significantly affected their access, including assessing the reasons for reduced benefits and how to address the issue
6. Whether and to what extent the project produced unintended (positive or negative) effects not captured in the project design or the evaluation design.

A process evaluation could furthermore:

1. Inform decisions on whether and how to proceed with a planned impact evaluation (i.e. used for evaluability purposes)
2. Inform the interpretation of impact evaluation findings and generalizability
3. Inform decisions about intervention refinement, scale-up and/or replication
4. Provide guidance on the design or implementation of future projects

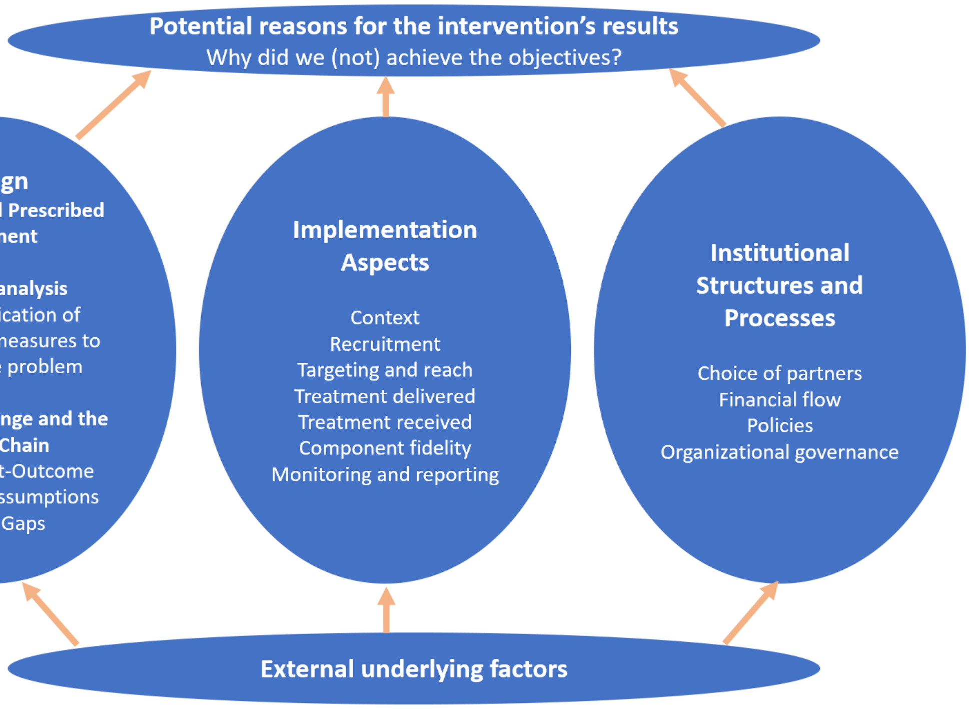
2. The four main dimensions of process evaluations

There are four main dimensions to process evaluations that may be particularly useful to apply to impact evaluations to identify what worked well for whom, when, where, how and why (or why not). These dimensions are illustrated in *Figure 2*.

1. The intervention design: Appropriateness of diagnosis and prescribed treatment
2. Implementation aspects[6]
3. Institutional structures and processes
4. The influence of external contextual factors beyond the control of project planners and managers

[6] Implementation aspects extend beyond implementation fidelity as outlined within the field of Implementation Research of Health interventions (see for example Peters et al. 2014 and Proctor et. al 2011). Here, we will focus on the elements of implementation that may affect the outcomes in any way (including the implementation's fidelity to the original plans/design/theory of change), because the main purpose and focus of a process evaluation is to identify the elements that have affected the intervention's results in a positive or negative way.

Figure 2: The process evaluation framework: Various elements affecting project outcomes


Potential reasons for the intervention's results
Why did we (not) achieve the objectives?
Design
Diagnosis and Prescribed Treatment
Problem analysis and identification of appropriate measures to address the problem
Theory of Change and the Logical Chain
Input-Output-Outcome
Underlying Assumptions
Logical Gaps
Implementation Aspects
Context
Recruitment
Targeting and reach
Treatment delivered
Treatment received
Component fidelity
Monitoring and reporting
Institutional Structures and Processes
Choice of partners
Financial flow
Policies
Organizational governance
External underlying factors

The first three dimensions (design/implementation/institutional framework) are influenced by the external context (economic, political, organizational, legal, socio-cultural, etc.) within which the project is implemented. These three dimensions and the external factors will be outlined in the following sections.

2.1 Intervention Design

Check the appropriateness of the diagnosis and the prescribed treatment.

Imagine you wake up one day with a headache, sore throat and fever. You drag yourself out of bed and head to your doctor to ask them for something to make you feel better. However, had you first looked up your symptoms on the Internet, you would have been surprised to find out that headache, sore throat and fever can be caused by 136 different conditions, among them typhoid fever, measles, brain tumor and COVID-19. Most probably, the doctor would prescribe common flu medication and you would feel better soon, but what if you had any of these other, more serious illnesses?[7]

Development interventions have similarities to medical treatments: If you treat superficial symptoms rather than the underlying pathology, or if you give the wrong medication, you will not cure the illness. In medicine, you would not say that the medication was ineffective in general; you would say that the doctor misdiagnosed the pathology. Similarly, in international development, we can only judge the effectiveness of an intervention after we have ascertained that it was designed to address the main

[7] The first three paragraphs of this section were first published on 3ie's blogpost in 2018, written by Marie Gaarder and Vibecke Dixon: Misdiagnosis and the evidence trap: a tale of inadequate program design | 3ie (3ieimpact.org)

underlying problem or *binding constraint*. Yet, all too often, as impact evaluators, we judge the effectiveness of development interventions without knowledge of whether policymakers and aid agencies established the correct diagnosis of the root cause (or causes) of a certain development problem and whether they designed an appropriate intervention to address it.[8]

Making the right diagnosis is a necessary condition to achieve impact in development interventions. As a theoretical argument, this is generally accepted. Nevertheless, often in practical impact evaluation work, scant attention is paid to this.

2.1.1 Problem Analysis

Checking the appropriateness of the *diagnosis* is, thus, of crucial importance in process and impact evaluation. We need to ask the question whether the correct underlying problem(s) was/were identified and sufficiently evidenced in the planning process. The appropriateness of the interventions' *problem analysis* is assessed by looking at the identified problem and its underlying alternative causes.[9]

For example, an education project in Mali had identified the problem *children not going to school*, because there were many children not enrolled in school and children not attending school despite being enrolled. The prescribed treatment was *awareness-raising activities for parents*. The underlying assumption was that the reason children are not enrolled or not attending school is that their parents are unaware of the importance of education.

Now, if the reasons for children's non-attendance or non-enrolment in school were any other than their parents' unawareness, the project's awareness-raising activities would not have addressed or solved the real problems. Alternative reasons for children's non-attendance/non-enrolment in school could potentially be:

- Conflict/violence/danger on the way to school (shooting/kidnapping, etc.)
- Gender issues (only male teachers/lack of appropriate sanitary facilities, etc.)
- Poverty (food security issues, children having to work to support their families, etc.)
- Climate change (the route to school is flooded parts of the year, etc.)

[8] No single project can or should try to solve all the levels of inter-related problems affecting the residents of a particular village or district. However efficient a cooking stove, it cannot solve all of the climate-related problems affecting the village, nor the national and international terms of trade that influence the prices villagers obtain for their produce. So, the assessment of the program design must determine whether the design is appropriate given the level of program resources and the local and regional factors affecting it. Also, is there a logical relationship between this and another program, program and policy activities underway and planned for the community and district. Finally, the evaluation must also assess whether due consideration has been given to the factors affecting the program's sustainability, and in many cases, whether it is likely to be replaced.

[9] There are many problem analysis tools and approaches—far more than can be covered in these guidelines. One example is root cause analysis, which many development organizations use. The Millennium Challenge Corporation (MCC) provides a good description of their approach to root cause analysis here: https://www.mcc.gov/resources/story/story-cdg-chapter-6-guidelines-for-root-cause-analysis , in addition to referring to other useful guidance resources.

If the reason the children's non-attendance in school in this project area is *not* that their parents are unaware of the importance of education but any of the alternative reasons, a project designed to raise parents' awareness is not going to result in a higher enrolment/attendance rate of children in school, because the project does not address the actual underlying reasons for the problem and the intervention will not achieve its objective (of increased school enrolment/attendance).

If an impact evaluation or a systematic review does not assess correctly *the diagnosis and the prescribed treatment* of a project, the wrong conclusion may be drawn—e.g. that *awareness-raising activities for parents do not work to increase children's attendance/enrolment in school*, when in fact the reason the objective was not reached (more children in school) might have been that the wrong underlying cause had been addressed. Awareness-raising activities for parents may or may not work to increase children's school attendance where the underlying cause is parents' unawareness, but we would not have any evidence of that under the circumstances.

It is, therefore, of crucial importance to do a solid problem analysis at the appraisal stage of a project. Moreover, any evaluation of the project should pay attention to and assess the original problem analysis underlying the design to examine the appropriateness of both the *diagnosis* and the *prescribed treatment*.

Many agencies tend to identify the problem they are able to address—a transport agency will identify the poor quality of roads or the lack of school transport as the reason for low school attendance, while the ministry of health might argue that malnutrition is the main cause of low attendance. So, in the assessment of the appropriateness of the diagnosis, another important aspect that is often overlooked concerns whose perspectives were taken into consideration when analyzing the problem and designing the solution. A transport economist might consider the design of a rural road project as appropriate to provide farmers with easier access to markets. However, did the diagnosis take into consideration the perspective of mothers with young children who are concerned about safety issues when a road goes through the village? Or the perspective of women who would have preferred the construction of a stairway and footbridge to make it easier and safer to herd their goats or reach their fields during the rainy season?

Another major issue for any project that involves major resettlement, such as roads, housing or irrigation, is whether its impacts on all affected groups have been considered. There are many projects where the views of landowners are considered, but the impacts on vulnerable groups such as the landless, indigenous groups or women farmers (compared to male farmers who are often considered the household representative) have not been addressed. Political pressures often come into play, because some government agencies or major power groups, such as large landowners, may intentionally ignore certain vulnerable groups because they do not wish to pay compensation to these groups.

The critical questions to assess the appropriateness of the diagnosis are:

1. Have the correct underlying problems been identified and sufficiently evidenced?
2. Have all affected populations been identified and their concerns addressed?
3. Is the *prescribed treatment* necessary and sufficient to address the underlying reasons for the problems, including the concerns of all affected populations?

2.1.2 The theory of change and the causal chain

Assessing the intervention's theory of change with the intention to detect any underlying assumptions and potential gaps in the logical chain is also important, because gaps in the logical chain or non-attention to underlying assumptions may be the cause of an intervention's failure to achieve its objective. Process evaluation results can be used to test theory (or parts of theory) of change, as well as to create a new one.

All development interventions are based on an explicit or implicit theory about how intended project outputs and impacts are to be achieved and the factors constraining or facilitating their achievement.[10] The theory of change (sometimes also called program theory) is an explicit theory or model of how the project is expected to yield the intended outcomes.[11] While program theory models can be used in all evaluations, they are particularly useful to help explain whether failure to achieve objectives is due to a faulty design or a faulty or ineffective implementation.[12]

Theories of change outline a sequence of events leading to outcomes—they explore the conditions and assumptions needed for the change to take place, make explicit the causal logic behind the project and map the project interventions along logical causal pathways. A theory of change can be modelled in different ways, using theoretical models, logical frameworks or results chains. All models include the basic elements of a theory of change: (1) a causal chain, (2) external conditions, and (3) underlying assumptions/logical gaps.

In the following paragraphs, we will outline the example theories of change in results chains, because results chains are simple and clear and the most often used to outline theories of change for development interventions. A results chain sets out the sequence of the intervention's inputs, activities and outputs that are expected to affect outcomes and objectives, which is useful for development interventions because it clarifies the causal logic and sequence of events in a project. Results chains are also useful for monitoring and evaluation, because they make evident what kind of information needs to be monitored and what outcome changes need to be included when the intervention is evaluated. In addition, they outline the intended inputs and activities in a way that they may be assessed to see whether they are necessary and sufficient to address the underlying causes of the problem (and thereby, achieve the objective).

A basic results chain will map the elements listed in *Table 1.*

[10] See also Bamberger, M et al. (2006:39) and Bamberger and Mabry (2020:25ff).

[11] See also Rogers, P et al. (2000:5).

[12] See for example Bamberger and Mabry (2020), Lipsey (1993), and Weiss (1997).

Table 1: Basic results chain[13]

Inputs	The financial, human, material, technological, and information resources used in the project.
Activities	Actions taken or work performed to convert inputs into outputs.
Outputs	The tangible goods and services project activities produce; They are directly under the control of the implementing agency.
Outcomes	The intended or achieved short- and medium-term effects of an intervention's outputs, usually requiring the collective effort of partners. Such results are likely to be achieved once the beneficiary population uses the project outputs. Outcomes represent changes in development conditions that occur between the completion of outputs and the achievement of impact. They are usually achieved in the short to medium term, and the implementing agency has little or no control over them.
Objectives/Impacts	The final project goals. Long-term economic, sociocultural, institutional, environmental, technological or other effects on identifiable populations or groups produced by a project, directly or indirectly, intended or unintended. They can be influenced by multiple factors and are typically achieved over a longer period of time.

The results chain has three main parts:

Table 2: The three parts of a results chain[14]

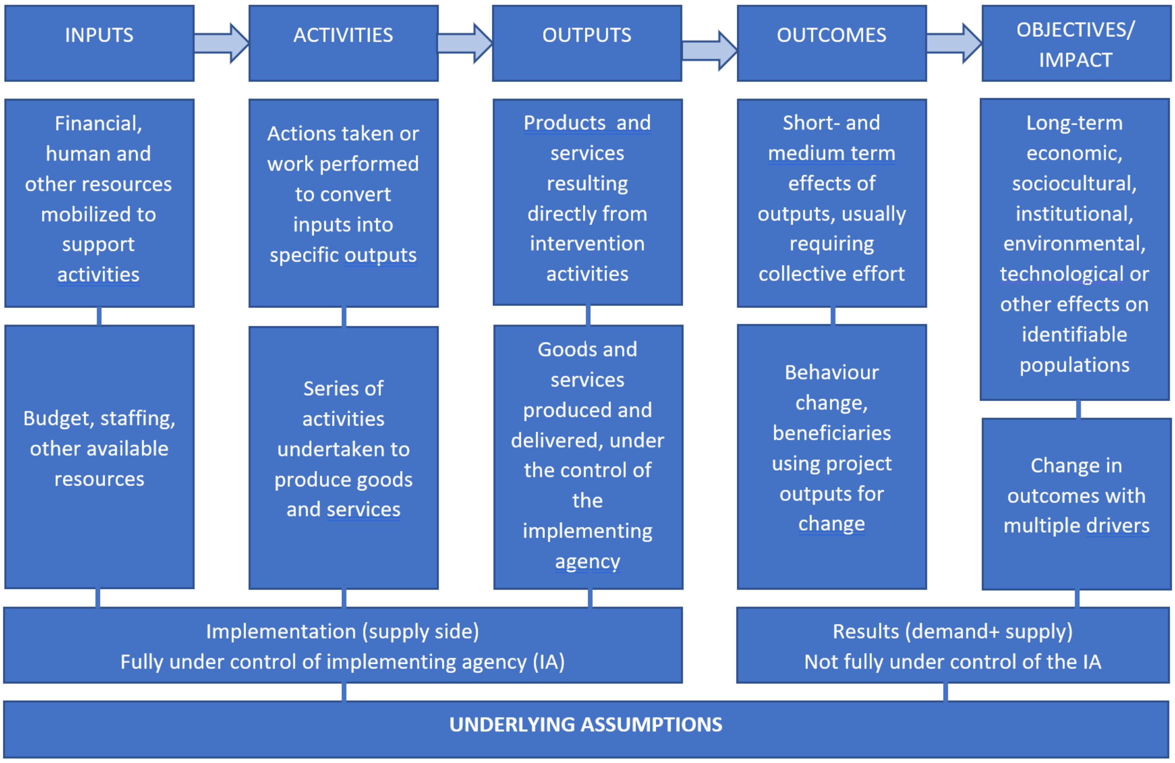
Implementation Inputs, Activities and Outputs	Planned work delivered by the intervention, including inputs, activities and outputs. These are the areas that are under the direct control of the project (implementing agency) and can be directly monitored to measure the intervention's performance.
Results Outcomes and Impact	Intended results consist of the outcomes and the final objective of the intervention, which are not under the direct control of the project and are contingent on behavioral changes by the intervention's beneficiaries. The project's outcomes and impact are answers to the question: *How have people's lives been changed (due to the project intervention)?*
Assumptions and Logical Gaps	These refer to conditions that are necessary for the intervention to work as intended. There may be logical gaps in the results chain where the underlying assumptions are not identified and tested—they have to do with logical links we would take for granted without testing their validity, such as for example, that the output *training of x number of people* will automatically lead to behavior change and *strengthened capacity*. It is then taken for granted that understanding, learning and uptake are happening automatically because people are exposed to new knowledge. They include evidence from literature or fieldwork on the proposed causal logic and the assumption on which it relies, references to the performance of similar interventions, and mention of risks that may affect the realization of intended results and any mitigation strategy put in place to manage those risks.

[13] Table 1 is based on the description of the results chain by Gertler et al. (2001:pp. 24–25), with some modifications for the purpose of these guidelines.

[14] Table 2 is based on the description of the three parts of the results chain by Gertler et al. (2011:pp. 24–25), with some modifications for the purpose of these guidelines.

A results chain may be drawn up as shown in *Figure 3*.

Figure 3: Outline of a theory of change in the form of a results chain[15]


INPUTS
ACTIVITIES
OUTPUTS
OUTCOMES
OBJECTIVES/ IMPACT
Financial, human and other resources mobilized to support activities
Actions taken or work performed to convert inputs into specific outputs
Products and services resulting directly from intervention activities
Short- and medium term effects of outputs, usually requiring collective effort
Long-term economic, sociocultural, institutional, environmental, technological or other effects on identifiable populations
Budget, staffing, other available resources
Series of activities undertaken to produce goods and services
Goods and services produced and delivered, under the control of the implementing agency
Behaviour change, beneficiaries using project outputs for change
Change in outcomes with multiple drivers
Implementation (supply side)
Fully under control of implementing agency (IA)
Results (demand+ supply)
Not fully under control of the IA
UNDERLYING ASSUMPTIONS

While the first three components of this model (inputs/activities and outputs) may be directly controllable by those managing the intervention, the outcomes, impacts and the intervention's sustainability[16] depend to a considerable degree on external factors over which the implementation agency usually has little or no control. It is important, however, that external factors are acknowledged and accounted for in the project design and implementation of the intervention. A good theory-based design will take into account competing theories as to how a project works to be able to capture unintended effects and other *surprises*.

The following outlines an example of what a theory of change for a specific intervention may look like: The ministry of education of a country is introducing a new approach to teaching mathematics in high school.[17] As shown in *Figure 4*, the inputs to the project would include staff from the ministry, high school teachers, a budget for the new mathematics project and the municipal facilities for teacher training. The project's activities consist of designing the new mathematics curriculum, developing a teacher

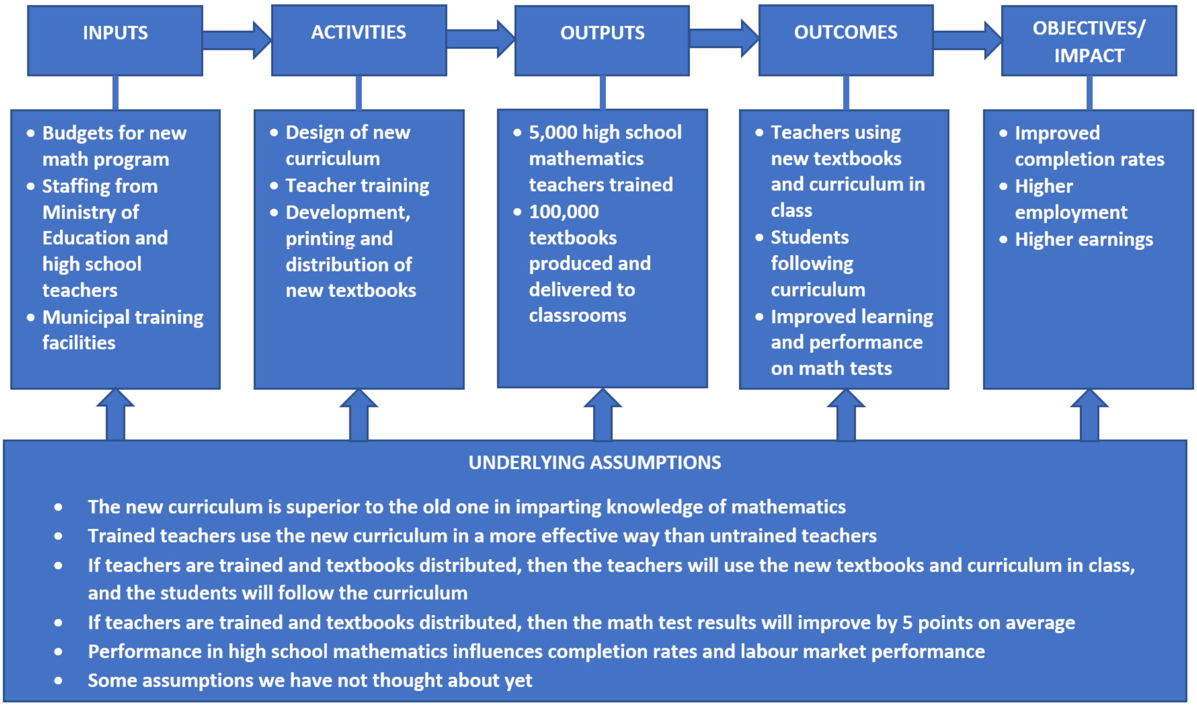
[15] This figure is based on Figure 2.1: *What is a Results Chain?* on page 25 in Gertler et al. (2011), with some modifications for the purpose of these guidelines.

[16] Some results chains include an extra step for sustainability (like we have done in Table 3 below). Sustainability is often the weakest point in program design, because it largely occurs after the project is completed and when the agency has little control over events. It is also difficult to monitor, because many sustainability events occur after monitoring and other data collection sources have ended. Many projects fail due to lack of attention to sustainability.

[17] This example is taken from Gertler et al. (2011:25–26).

training project, training the teachers, and commissioning, printing and distributing new textbooks. The outputs are the number of teachers trained, the number of textbooks delivered to classrooms and the adaptation of standardized tests to the new curriculum. The short-term outcomes consist of teachers' use of the new methods and textbooks in their classrooms and their application of the new tests. The medium-term outcomes are improvements in students' performance on the standardized mathematics tests. The impacts are increased high school completion rates, and higher employment rates and higher earnings for graduates.

Figure 4: Results chain for a high school mathematics project[18]



INPUTS

- Budgets for new math program
- Staffing from Ministry of Education and high school teachers
- Municipal training facilities

ACTIVITIES

- Design of new curriculum
- Teacher training
- Development, printing and distribution of new textbooks

OUTPUTS

- 5,000 high school mathematics teachers trained
- 100,000 textbooks produced and delivered to classrooms

OUTCOMES

- Teachers using new textbooks and curriculum in class
- Students following curriculum
- Improved learning and performance on math tests

OBJECTIVES/ IMPACT

- Improved completion rates
- Higher employment
- Higher earnings

UNDERLYING ASSUMPTIONS

- The new curriculum is superior to the old one in imparting knowledge of mathematics
- Trained teachers use the new curriculum in a more effective way than untrained teachers
- If teachers are trained and textbooks distributed, then the teachers will use the new textbooks and curriculum in class, and the students will follow the curriculum
- If teachers are trained and textbooks distributed, then the math test results will improve by 5 points on average
- Performance in high school mathematics influences completion rates and labour market performance
- Some assumptions we have not thought about yet

2.1.3 Underlying assumptions and logical gaps

Taking external factors and context into account facilitates the identification of critical underlying assumptions. A key element of project theory models is the identification and monitoring of critical (underlying) assumptions about the causal links between the different stages of the theory of change model.

A weakness of many theories of change is that they do not spell out the critical assumptions that should be tested at each stage of the model. A well-formulated set of assumptions is the building block (logical links) in a project design that is essential to the effective implementation. However, many theories of change and results frameworks only include general assumptions that do not articulate these logical links—e.g. *assuming the government continues to provide funding*. A key element of a good theory of change is that it identifies the logical links and gaps underlying the project design, and then assesses their validity as the project evolves. For example, if a women's microcredit project includes training on setting up an accounting system for the business, there is an

[18] This figure is based on figure 2.2 on page 26 in Gertler et al. (2011), with some modifications for the purpose of these guidelines.

implied logical link stating that lack of accounting skills is one of the barriers to launching a successful business. Or it may be assumed that encouraging men and women to plan the project together (whereas previously, they planned it separately) will make the project more successful and responsive to the specific needs of women participants. These implicit logical links may be spelled out and then periodically tested throughout the project.

Imagine a project with the objective to strengthen women's economic and social empowerment, where the provision of credit for women is the major input. This is based on several critical underlying assumptions—lack of access to credit is one of the main constraints for women to start small businesses and if women have access to credit, this will significantly increase the number of businesses they will start. It is also assumed that women will be able to control their own income, because an increase in women's income will automatically lead to their economic empowerment. In *Table 3*, some of these critical assumptions are outlined according to how they may appear in the results chain.

Table 3: Critical assumptions to be tested at different stages of a project

Stage of project	Critical assumption to be tested
Design	Poor women have the skills needed to operate viable income-generating projects, but lack only capital.
	Women are able to decide for themselves what businesses to start/expand.
	Women will be able to control how the loan is used, and the money will not be appropriated by the husband.
Inputs	Access to credit, in a form that women can control, is critical to enhancing women's access to economic opportunities.
Implementation process	The creation of solidarity groups through which loans are approved and technical support provided is essential to enable women to control the use of their loans and manage their small businesses.
	Solidarity groups will be able to select their own members without any outside pressure.
Outputs	Women will use loans to invest in small businesses (not just to pay off depts or pay for consumption or ceremonial activities).
	Women will be able to control the use of the loan (despite cultural traditions that economic resources are controlled by male household members).
Outcomes	If women produce goods, they will be able to market them.
	Women's businesses will be profitable.
	Women will control or share in the control of profits.
Impact	Profits will increase household consumption, women's savings and quality of life of members of their households.
Sustainability	Women's solidarity groups will be able to continue providing loans after the project's external credit and support has ended
	Women's businesses will continue to operate and grow.

In a similar project, where the startup of women's kitchen gardens in Zambia were to enhance women's economic empowerment, the project looked successful on paper—the participating women had produced and sold vegetables, which had earned them a significant amount of money at the end of the season. The initial conclusion was that "this kitchen garden project has increased participating women's income significantly and has, thus, contributed to women's economic empowerment." However, when one of the authors visited the project with a study team, the women were very adamant that they did not want to participate when the project was to start up again, the following season, saying “We will never do anything like that again!” When asked why, they explained that their husbands had taken all their money and bought new wives with their hard-earned income.

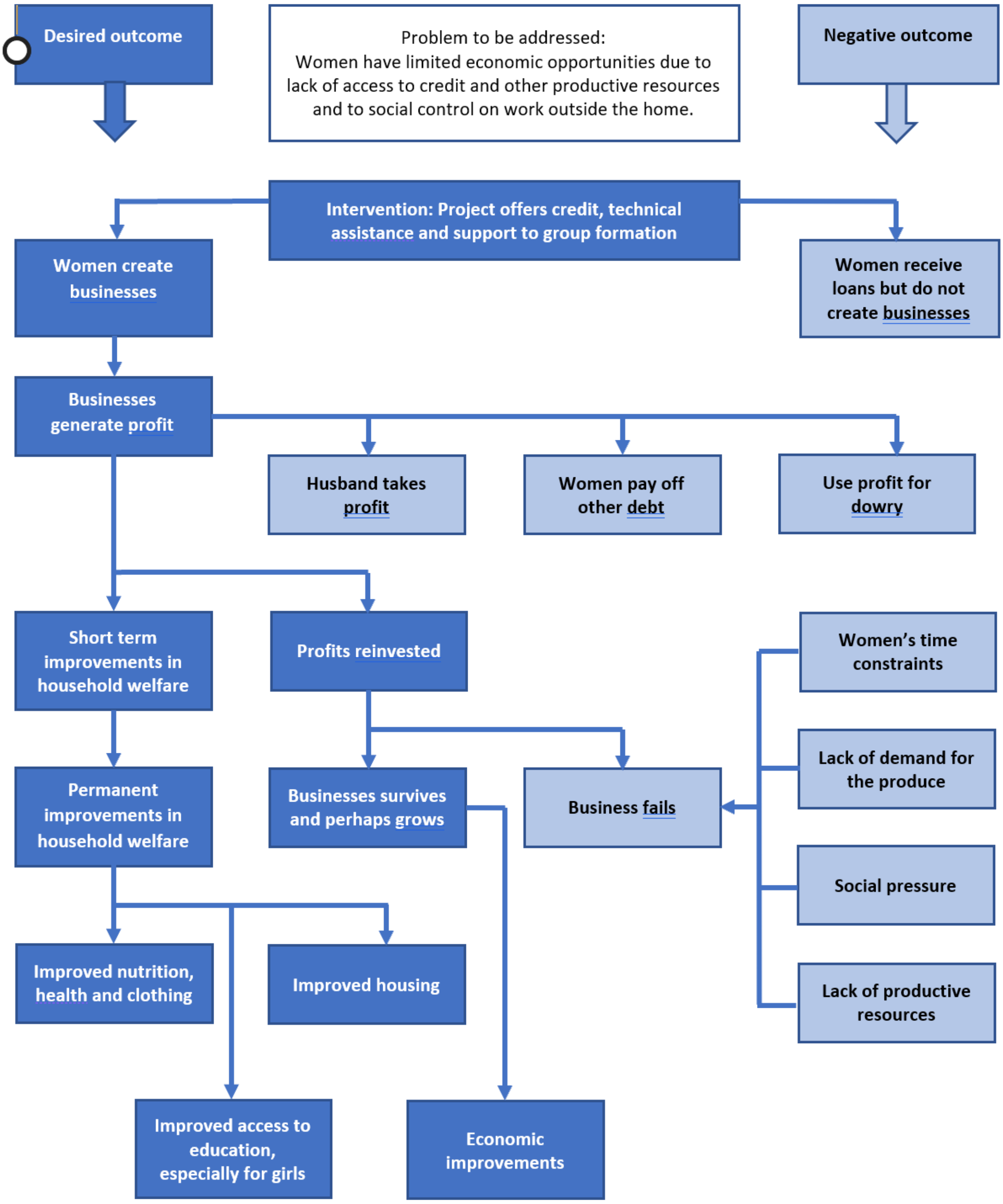
This is a typical example of a development intervention where the sociocultural context, especially the gender and power relations, had not been taken into account in the design and the underlying critical assumption *increasing women’s income will automatically lead to women’s economic empowerment* had not been identified, problematized or tested.

Similarly, the Women’s Goat Scheme in Zambia was for a long time one of the flagship projects of an international development organization. It had been designed by highly professional agricultural economists at the organization’s headquarters in Europe. They knew that in most parts of Africa, men would work with large animals and women with smaller ruminants. The underlying assumption of the project was that a goat scheme would automatically target female beneficiaries. However, when visiting the project a couple of years into the implementation process, we found that only about 6 percent of the project beneficiaries were women and 94 percent were men. This was because in that specific location of Zambia, only men worked with goats due to traditional beliefs that it was taboo for women to be near goats. While the project was up and running and the beneficiaries did relatively well with the goat rearing, the project had failed in targeting and supporting women, and thus, did not reach its objective of increasing women’s income.

These examples show how crucial it is to know the context, and identify and address critical underlying assumptions both in design and in evaluation.

Figure 5 on the following page shows how the critical assumptions in *Table 3* may be expressed in the form of a results chain diagram. The left-hand side of the figure shows the intended chain of events—women use the loans to create businesses that generate profits that contribute to improvements in household welfare and are reinvested to ensure growth and sustainability of the business. The right-hand side of the figure (in a lighter shade of blue) identifies the different reasons the project might fail to achieve its objectives—women do not use the loans to create businesses, the profits are taken by the husband or used to pay off debts or provide dowries, or the businesses fail for other reasons, such as time constraints and social pressure.

Figure 5: A results chain model of the women's microcredit project[19]


Desired outcome
Problem to be addressed: Women have limited economic opportunities due to lack of access to credit and other productive resources and to social control on work outside the home.
Negative outcome
Intervention: Project offers credit, technical assistance and support to group formation
Women create businesses
Women receive loans but do not create businesses
Businesses generate profit
Husband takes profit
Women pay off other debt
Use profit for dowry
Short term improvements in household welfare
Profits reinvested
Women's time constraints
Lack of demand for the produce
Permanent improvements in household welfare
Businesses survives and perhaps grows
Business fails
Social pressure
Improved nutrition, health and clothing
Improved housing
Lack of productive resources
Improved access to education, especially for girls
Economic improvements

2.1.4 Reconstruction of theory of change

Not all project designs or plans have an explicit, written theory of change. In those cases, it will be necessary for the evaluators to re-construct the theory of change by asking questions such as: What were the project's intentions? What was the identified problem and what were its underlying assumptions? What was causing the problem? Did the project design address the root causes of the problem? What were the main objectives

[19] *Figure 5* is based on Figure 10.4 on page 161 in Bamberger and Mabry (2020), with some modifications for the purpose of these guidelines.

of the intervention? What were the inputs? What were the outputs inputs were expected to lead to? And in turn, what outcomes were these outputs expected to lead to? Were the components/interventions appropriate (sufficient and necessary) to address the problem? Were there any underlying assumptions or logical gaps in the theory of change? Was the logical chain clear and convincing? Can measurable results be attributed to the project's intervention?

If possible, the reconstructed theory of change should be checked with those who designed the project. Although they may not have drawn up a theory of change, they would most likely have useful inputs to a reconstructed one.

One of the challenges of reconstructing a theory of change is that the process is often conducted by consultants with only limited input from project management and perhaps other stakeholders. Senior staff are often too busy to spend sufficient time and frequently agree with the proposals from the consultants. In contrast, when a new project is planned (and funding is still negotiated) key staff are more willing to be involved. In some cases, staff may also be reluctant to participate in the discussions, because they perceive theories of change to be difficult to understand.

2.1.5 Relevant evaluation questions to be considered

The following list is not intended to be perceived as a complete list of relevant questions for process evaluations. The questions are meant as suggestions evaluators might want to include in their evaluation design and to spark further ideas.

- Have the underlying reasons for the problem been identified and is there sufficient evidence to support them?
- Have all affected populations been identified and their concerns addressed?
- Is the suggested solution (project intervention) necessary and appropriate to address the reasons for the problem? Is supporting evidence of this presented?
- Is the theoretical framework (theory of change, results chain, process tracing[20]) clear and convincing?
- Are there logical gaps in the theory of change results chain?
- Are there any undetected underlying assumptions or logical gaps in the theory of change?
- Who was involved in the design or validation of the theory of change? Was there a participatory process or was the theory of change developed (and/or reconstructed) by a consultant?
- Have all relevant contextual factors been taken into consideration?
- Are the project components/project activities necessary and sufficient to achieve the objective, i.e. to solve the problem and address the factors that are causing the problem?

Below, we have inserted these evaluation questions into a table and linked them to relevant tools for evaluators to consider referencing where descriptions of the tools may be found.

[20] Beach, D and Pedersen, R, 2019. *Process-tracing methods*.

Table 4: Examples of research tools that can be used to address the different evaluation questions on design aspects (diagnostics, theory of change and project design)

Question	Tools
Problem identification: Is there sufficient evidence to support the identified problem? Have the underlying reasons for the problem been identified and is there sufficient evidence to support them?	Assessment of the intervention's problem analysis Literature review of academic literature (ethnographies, economic studies, etc.) and previous evaluations from the project area to provide relevant information regarding the problem area and potential underlying reasons for the problem. Millennium Challenge Corporation's Problem Analysis tool Key informant interviews Field visits for direct observation and participant observation, interviews and focus groups
Have all affected populations been identified and their concerns addressed?	Literature review Building questions into the quantitative impact surveys Participatory group consultations (Participatory Rural Appraisals (PRA), Rapid Rural Appraisals, (RRA) etc.), including the construction of social maps where the group can identify the characteristics of each household, including whether they are involved in the project Satellite images can visually identify whether there are any geographical groups not included in the project
Is the suggested solution (project intervention) necessary and appropriate to address the reasons for the problem? Is supporting evidence of this presented?	Analysis of project design Academic and evaluative literature Focus groups and key informant interviews
Is the theory of change (theoretical framework) clear and convincing? Are there any undetected underlying assumptions or logical gaps in the theory of change?	Analysis of the theory of change Key informant interviews Field visits/observations/interviews
Who was involved in the design or validation of the theory of change? Was there a participatory process or was the theory of change developed (and/or reconstructed) by a consultant?	Key informant interviews Project documents
Have all relevant contextual factors been taken into account?	Academic literature (ethnographies/economic studies/environmental/agricultural surveys, etc.) and previous evaluations Construction of complexity map
Are the project components/project activities necessary and sufficient to achieve the objective (i.e. to solve the problem and address the factors that are causing the problem)? Are the limits of the target population clearly defined? Is it clear who is eligible? Are the kinds of project outcomes clearly	Analysis of project design in light of contextual and sectoral knowledge (from literature review) Group consultation with the target population Review of project documents and interviews with project staff and other key informants

Question	Tools
defined? For example, are they only economic or do they include social and cultural changes, health and education, organizational changes or climate change? Is the time horizon over which changes are to be measured clearly defined?	

2.2 Implementation aspects

Check whether the implementation protocol for the prescribed treatments was complied with.

This dimension of the process evaluation assesses relevant aspects of the project implementation, and looks at how and to what degree the intervention was implemented as intended. The main question to be addressed is: How were the different components of the project implemented and how closely did this conform to the project plan, operations' manual or relevant sectoral good practice standards?

Process evaluation of implementation factors involves documenting and describing specific project activities—how much of what, for whom, when and by whom. Part of examining an intervention's implementation is looking at whether specific elements, such as facilities, staff, space or services, are being provided and established according to the plan. It includes monitoring the frequency and extent of implementation of selected elements[21] again with the focus on determining how these may have affected the intervention results.

There are many development interventions that do not have a clearly outlined implementation strategy with a high level of detail. In fact, there is a continuum from interventions, such as infrastructure, that usually have detailed and precise implementation plans to those with a strong community participation component, where implementation strategies are expected to evolve in cooperation with the community. In addition, in many interventions the objectives and implementation strategies for the infrastructure and other technical components may be clearly defined, but there is much less clarity with respect to social components, such as inclusion, gender equality, and so on. Also, certain projects address issues that are inherently complex and where implementation approaches gradually evolve, such as climate change.

There are also situations where there is no clear documentation of the implementation strategy and the evaluator may have to reconstruct the strategy based on interviews, relevant agency documents and in some cases, direct observation. This is similar to the process for reconstructing the theory of change, discussed earlier.

Once the methodology for tracking and evaluating the implementation process is defined, it might be useful to conduct an evaluability analysis to assess whether it would be possible to conduct the proposed evaluation, and whether it would provide the kinds of information needed to assess the adequacy of the implementation strategy and process.

[21] See Windsor et al. (1984:3) for a further outline of this.

Some of these elements explored here and as applied in a process evaluation may overlap with certain elements of Implementation Research, which is a growing research field that began in the health sector and is now being more widely applied. See footnote 22 for references to useful Implementation Research guidelines with elements that may

[22] See for example Peters et al. (2014) and Proctor et al. (2011). Implementation Research aims to cover a wide set of research questions, implementation outcome variables, factors affecting implementation and implementation strategies beyond our focus of assessing the implementation factors that might affect the intervention's outcome. Implementation Research has been defined as "the scientific inquiry into questions concerning implementation—the act of carrying an intention into effect, which in health research can be policies, programmes, or individual practices (collectively called interventions)." (Peter's et al. 2014). Implementation outcomes include acceptability, adoption, appropriateness, feasibility, fidelity, implementation cost, coverage and sustainability. These variables are used to assess how well implementation has occurred or provide insights about how this contributes to one's health status or other important health outcomes.
One of the earlier examples of a process evaluation that did look into implementation fidelity in a systematic manner was a public health process evaluation (CATCH) that, in the early 1990s, focused on four main areas: (1) participation—did teachers, food service personnel and public education specialists attend the training sessions?, (2) dos—were prescribed components of the program implemented?, (3) design fidelity—were the prescribed intervention components implemented according to protocol?, and (4) compatibility—did the CATCH programs fit the context of the schools, as well as the needs, expectations and values of the staff members and teachers? (See Linnan and Steckler 2002).
Baranowski and Stables (2000) work within the same concept framework and have listed the following 11 components of process evaluation: (1) Recruitment: attracting agencies, implementers or potential participants for corresponding parts of the program; (2) Maintenance: keeping participants involved in the programmatic and data collection; (3) Context: aspects of the environment of an intervention; (4) Resources: the materials or characteristics of agencies, implementers or participants necessary to attain project goals; (5) Implementation (fidelity): the extent to which the program is implemented as designed;
(6) Reach: the extent to which the program contacts or is received by the targeted group; (7) Barrier: problems encountered in reaching participants; (8) Exposure: the extent to which participants view or read (and understand) the materials that reach them; (9) Initial use: the extent to which a participant conducts activities specified in the materials; (10) Continued use: the extent to which a participant continues to do any of the activities; and (11) Contamination: the extent to which participants receive interventions from outside the program and the extent to which the control group received the treatment. This list provides a useful beginning framework for organizing conceptual thinking about process evaluation and developing consistent definitions to be used in the measurement of key process evaluation components.
Linnan and Steckler 2002:12 present the following components and definitions: (1) Context: aspects of the larger social, political and economic environment that may influence intervention implementation; (2) Reach: the proportion of intended target audience that participates in an intervention. If there are multiple interventions, then it is the proportion that participates in each intervention or component. It is often measured by attendance; (3) Dose delivered: the number or amount of intended units of each intervention or each component delivered or provided; dose delivered is a function of efforts of the intervention providers; (4) Dose received: the extent to which participants actively engage with, interact with, are receptive to and/or use materials or recommended resources; dose received is a characteristic of the target audience and it assesses the extent of engagement of participants with the intervention; (5) Design fidelity: the extent to which the intervention was delivered as planned. It represents the quality and integrity of the intervention as conceived by the developers. Fidelity is a function of the intervention providers; (6) Implementation: a composite score that indicates the extent to which the intervention was implemented and received by the intended audience; and (7) Recruitment: procedures used to approach and attract participants; recruitment often occurs at the individual and organizational/community levels.

be beneficially applied to process evaluation. However, it is worth noting that Implementation Research aims to cover a wide set of research questions, implementation outcome variables, factors affecting implementation and implementation strategies far beyond our narrower focus of *finding what has caused the intervention results*.

To improve and sustain successful development interventions, we need to identify the key components of an intervention that are effective, for whom the intervention is effective, and under what conditions it is effective. Part of this is to identify the extent to which all intervention components were actually implemented. This also includes assessing the quality and accuracy of the intervention as delivered to project participants.

Very few (if any) development interventions are implemented exactly as designed. Unexpected events often occur in real life and they may be difficult to plan for in an intervention design. It is, therefore, important for the evaluators to identify the *necessary conditions* for the intervention to yield the expected outcomes, and then assess how and to what degree these necessary conditions were met. There is a close link between the *necessary conditions* and the *underlying* or *critical assumptions* as presented earlier. For example, in a Conditional Cash Transfer project in Honduras, no difference in effects was found between villages that only received family cash and those that received both family cash and extra supplies to the schools and health centers. A necessary condition for the project to show different effects in different villages was that the schools and health centers had actually received the expected supplies, which in this case, they had not due to procurement issues. Consequently, the necessary conditions were not met.

There are often unexpected elements and unaccounted occurrences that affect project implementation in such a way that results are not achieved as expected. Frequently, while assessing the implementation aspects of an intervention, relevant contextual factors overlooked in the design become apparent. The evaluative process of finding these contextual and underlying factors bears similarities to detective work in that the aim is to uncover the unknown and detect the unexpected that has had a bearing on the implementation process and project results.

In their article "Avoiding Type III Errors in Health Education Programme Evaluations: A Case Study," Basch and colleagues (1985) introduced the concept of *Type III Errors*, which aligns well with looking at implementation fidelity. While researchers are familiar with Type I errors (e.g. rejecting a *true* null hypothesis, or a Type II error (e.g. failing to reject a *false* null hypothesis), a Type III error would ensue from "evaluating a program that has not been adequately implemented" (1985:316). Basch and colleagues argue that in addition to asking "did the project work," evaluators would need to ask whether the project was carried out as planned, and if not, how it varied from the original plan. This means that when interpreting monitoring data and other project results, evaluators must explore whether participants received and used what was delivered.

The Type III errors are a critical, but frequently overlooked element of impact evaluations. Many, but certainly not all, impact evaluations focus on assessing whether there is a statistically significant difference in outcomes between the project and control/comparison groups. Many of these evaluations do not assess implementation fidelity (Did all of the clinics receive and distribute the malaria tablets and the bed nets, as well as provided advice on how to use the treatment? Did all students receive the new teaching materials and were the teachers actually teaching regularly?). Consequently, the evaluation design implicitly assumes a high level of implementation fidelity and the lack of statistically significant outcomes is attributed to design failure. It is surprising how many impact evaluations do not systematically address how well the project was actually implemented. Also, if implementation issues are addressed, the analysis relies exclusively on monitoring and other reports produced by the project management, which tend to downplay any implementation failures that would reflect badly on the project.

While Type III errors are important when they reflect poor project management, they are even more critical if they reflect systematic (intentional or unintentional) biases. Examples of systematic biases might include low proportions of female-headed households or under-representation of women in community groups managing the project, under-representation of certain religious or ethnic groups or refugees, or political influence in the selection of project beneficiaries. Many of these factors are difficult to detect and require a systematic research focus, and normally, the information would not be found in project reports prepared for the funding agency.

The study of the Bangladesh Integrated Nutrition Project may serve as a good example to shed light on some of the most crucial implementation fidelity elements to look out for. The aim of the project was to improve mothers' and children's health through both nutritional counselling to mothers of young children and supplementary feeding (White and Masset 2006; White 2010b). The project adopted a growth-monitoring approach. This required monthly weighing of children from birth to 24 months and, if the child was severely underweight or its growth had faltered, enrolling their mothers into nutritional counselling sessions. However, the anthropological literature pointed to the widespread existence of joint families and the limited say in decision-making of women living with their mothers-in-law (White 1992).

As detailed below, this project's impact was undermined by both design and implementation failures. On the design front, a lack of attention to relevant sociocultural factors led to a failure to provide information about the project to mothers-in-law and husbands. This weakened the impact of nutritional counselling given to the young mothers, because they were not the key decision makers in obtaining and using food. Their husbands would be the ones to do the grocery shopping and their mothers-in-law would be in charge of the cooking (what to cook and how). Furthermore, the effectiveness of supplementary feeding was reduced by mistargeting, resulting in part from the inability of the village fieldworkers to read the growth charts intended to screen children for entry into the project. There were several underlying assumptions along the causal chain through which the project may have been expected to have positive impact on nutritional outcomes.

The following outlines a number of factors to look for when assessing the implementation aspects of a development intervention:

2.2.1 Do people know about the project and do they participate?

The first issue is *whether people indeed know about the project and participate*. Many development projects fall at the first hurdle because insufficient effort is made to explain the intervention to intended beneficiaries or make a realistic assessment of the relative cost and benefits for beneficiaries. This may be called acceptability, recruitment or reach in the different taxonomies referred to in footnote 20.

It is also important to assess whether the information people receive is correct. Many people may, intentionally or unintentionally, spread incorrect information. Some common kinds of misinformation are:

- Who is eligible to participate? For example, as White (2013, Chapter 6) pointed out, for the Self-Help Group (SHG) livelihoods project in India, the family unit was considered the target beneficiary, while in fact, the individual women should have been targeted (i.e. allowing two women from the same household to participate).
- How much does it cost to participate (when in fact it is free)?
- How effective or safe is the project?
- How to enroll?

This analysis is important because project management assumes that everyone will have received and understood the message conveyed, and used the material put out.

2.2.2 Are the right people targeted?

Second, *the people targeted have to be the right ones*. Who benefits from the project? Targeting analysis is important in both process and impact evaluations, and should be carried out at different levels.

For example, in the case of social funds, it was found that the use of poverty maps means that social funds in many countries focused on the poorest districts, but within those districts, it was the better off communities that were more likely to access project resources (World Bank 2002). Conversely, in the case of rural electrification, better off communities were more likely to connect, but poorer households in connected communities remained unconnected for many years due to their inability to afford the connection charge (World Bank 2008).

In the Bangladeshi nutrition project example, the right children had to be admitted to the project (i.e. those who were growth faltering or malnourished). Data showed substantial mistargeting with both Type I (children not being in the project when they should be) and Type II errors (children who should not be in the project being enrolled). This was due to nutrition practitioners' lack of understanding of how to use and read the growth charts. This mattered a great deal for project impact, because the most malnourished children did benefit, so average impact would have been greater had the project concentrated on such children, while in fact, resources were going to children who did not benefit.

The process of targeting can also be culturally sensitive. For example, in Indonesia, a number of agencies wished to target the poorest households, either using quantitative survey data or asking the community to use techniques such as social maps. However, it was found that many Muslim communities objected to dividing the community into poor and non-poor groups. One of the responsibilities of the local mosque was to maintain close contact with the community and identify individuals or households who were in need of assistance at a particular point in time, and for a variety of reasons, not just poverty. It has also been found that in various regions, poverty is considered to be shameful and respondents were not willing to identify their neighbors as poor.

2.2.3 Has the treatment been correctly applied (dose delivered)?

This has to do with the number or amount of intended units of each intervention or each component delivered or provided. Dose delivered is a function of efforts of the intervention providers. Have the pregnant women attended all check-up sessions? Have the children received all vaccine doses? Did the school children receive the learning materials?

2.2.4 Is the treatment taken by the right beneficiaries (targeting)?

For the supplementary feeding to have a beneficial impact, it has to be *supplementary*. However, in our Bangladeshi example, there was both *leakage*—the food was given to someone other than the person it was intended for, this was particularly the case for the supplement given to pregnant women, and *substitution*—the food was taken in place of a meal that would otherwise have been given, and this was particularly the case in the poorer households. The project impact was thus undermined by weak and missing links in the causal chain. This is another example where Type III errors would only be detected through close observation, because families were aware that the project had strict requirements about who should receive, for example, nutritional supplements, so under-nourished children or pregnant women would try to conceal the fact that they shared their supplement with other household members.

2.2.5 Does exposure lead to uptake (dose received)?

Perhaps one of the more common critical underlying assumptions in project theories of change is the assumption that training or introducing new knowledge automatically leads to behavior change. Evidence from several studies suggest that this is not always the case. It is, therefore, important to examine whether:

1. People who were exposed to training did understand and learn
2. They put the new knowledge into practice

Uptake or *dose received* covers the extent to which participants actively engage with, interact with, are receptive to and/or use materials or recommended resources. This assesses the extent of participants' engagement with the intervention.

The Bangladeshi nutrition study showed that mothers acquired the knowledge conveyed to them, but many did not put it into practice for contextual power reasons: their husbands and mothers-in law made the decisions, not the young mothers. Community nutrition practitioners had learned how to conduct weighing sessions, but not how to interpret a growth chart, which lead to mistargeting. The first is an example of where understanding has taken place (presumably), but uptake is hindered by contextual cultural power relations, and the second is an example where the training failed to provide appropriate understanding.

In a Norwegian-funded community schools for girls project in the Federally Administered Tribal Areas in Pakistan, monitoring data showed very high attendance rates for both girls and boys in schools and this was taken as an indication of the project's success, partly because a different pattern would often appear in other Pakistani schools where enrolment rates for boys and girls could be fairly equal, but girls' attendance rates would be low. When visiting the schools for a Norad evaluation, we asked to see the attendance records and found, to our surprise, that they were all—in all the schools we visited—filled in months ahead of time with *P* for present for all the students. The teachers explained that they had learned to fill in attendance records during their two-weeks' training course to become teachers. They made sure to fill in the records in advance, because they saw that as an important part of their job and reporting. They had not been taught or had not understood the purpose of attendance keeping, which in this case, led to incorrect and useless monitoring data. This is an example of failure in uptake, because the teachers had not understood the training they received and had not put the new knowledge into appropriate and meaningful practice.

2.2.6 Is monitoring data credible and reliable?

The example above regarding the attendance records is also an example of how lack of uptake may lead to incorrect and useless monitoring data. Similarly, as reported by White (2008, 2015), there were discrepancies in the monitoring data in a project encouraging women-only grassroots organizations at the village level called Self-Help Groups (SHGs) in the Indian state of Andhra Pradesh. Women joining these groups were required to make a monthly contribution of Rs.30, that is a little under $1. These contributions funded revolving loans. By 2007, more than 700,000 such groups had been formed, partly facilitated by two projects supported by the UK Department for International Development (DFID) and the World Bank, which provided funds and technical training to SHGs. The World Bank's Independent Evaluation Groups' evaluation of these projects used panel data collected in 2005 and 2007. Responses to the village questionnaire, which listed all the SHGs in the village, confirmed a continued rise in the number of SHGs and even a small rise in the average size of the groups over this two-year period. But the individual-level data (i.e. interviews with villagers) showed a drop in participation in SHGs from 42 percent of all eligible women in 2005 to just 30 percent in 2007. There was a discrepancy between the village-level data, which showed SHG membership to be rising, and the individual data, which showed it to be falling. This apparent discrepancy was readily explained by the qualitative data that the evaluation team collected alongside the quantitative survey. When villagers were asked about the number of SHGs in their village, the response had almost invariably been along the lines of, 'there are 22 SHGs of which 7 are currently not functioning." It showed that once started, an SHG stays on the books even if it ceases to function, and this was inflating the village-level (and state-level) estimates of the number of operational SHGs.

Save the Children Norway were supporting several well-functioning education projects, but were surprised to find that in some countries, monitoring data showed a sharp decline in children's reading abilities from 2nd to 5th grade after project implementation, suggesting that the project had totally failed (in fact, had had a reverse effect) on children's reading and learning abilities. Upon further examination, it was found that the reading tests done in grade 2 were in the children's mother tongue, while the reading tests for the 5th graders were in English. The 5th grader's reading test did not test the children's reading ability, but their English skills and results from the two tests were not comparable.

Beneficiaries' perception studies may also be biased, depending on how they were executed, who conducted them (project staff or independent evaluators), how the questions were phrased, how the informants were selected and whether the informants' interests (in e.g. being selected for the next intervention) might have influenced their responses. Also, beneficiaries' perceptions might not always be accurate with regard to the actual quality of services rendered. Some doctors might receive a high score by patients because they are perceived to be attentive and considerate (which are positive traits for a doctor), but it says nothing about the quality of care given, which might be assessed through direct observation, e.g. whether the doctors are asking the appropriate questions to make a diagnosis.

There is also the question of whether the right kinds of data are collected. Often, the required data are not available or are difficult to measure, and many agencies rely on whatever data are easily available even if they do not capture the required indicators. For example, it is very difficult to collect data on gender-based violence, so an agency may rely on claims registered with the local police station, which could provide a misleading picture. The question of construct validity must be assessed. Another example might be using the number of gender violence awareness classes as an indicator of an effective project to reduce gender-based violence, when there may be no evidence that the projects have any effect.

2.2.7 What to do when surprises arise and the unknown is uncovered?

When the unknown is uncovered and surprises arise, depending on what is found, this might need to be explored further. If deviations are found, like in the example with the kitchen gardens in Zambia where the husbands took their wives' income to buy new wives, the evaluation might want to further explore the frequency of the deviation, in addition to assessing the intervention's sustainability. It is important to have a flexible evaluation design so that surprise findings can be incorporated into the study when needed.

2.2.8 Relevant evaluation questions to be considered

The following list is not intended to be perceived as a complete list of relevant questions for process evaluations. The questions are meant as suggestions evaluators might want to include in their evaluation design and to spark further ideas.

- How were the different components of the project implemented and how closely did this conform to the project plan,[23] operations' manual, or relevant sectoral good practice standards?
- What was the originally suggested process and implementation plan?
- What did the actual process and implementation look like?
- Were all intervention components implemented as planned?
- What deviations were there from the plan?
- What were the reasons for the deviations?
- How did these deviations affect implementation and results?
- To what degree was the quality of the services acceptable/up to standard/high quality?
- Who had access to and/or used the services and who did not? Why did certain groups (not) use the services (reach)?
- Were the right people targeted? Were a sufficient number of beneficiaries exposed to the intervention and were the appropriate beneficiaries targeted (targeting)?

[23] As mentioned earlier, not all development interventions have clear designs or implementation plans. In those cases, it will be necessary for the evaluators to recreate the intervention's theory of change, and analyze the degree to which the project activities are necessary and sufficient to achieve the objectives.

- Was the treatment appropriately applied (dose delivered)?
- Was the treatment taken by the right beneficiaries (uptake)?
- Did exposure lead to uptake (were people using the new knowledge gained)?
- Were monitoring data credible and reliable?
- What are the external and contextual factors that may have materially altered the nature or strength of the intervention and its results?

Below, we have inserted the most relevant evaluation questions into a table and linked them to relevant tools for evaluators to consider referencing to where descriptions of the tools may be found.

Table 5: Examples of research tools that can be used to address the different evaluation questions on implementation aspects

Question	Tools
Evaluation fidelity: How closely did project implementation comply with the project protocol?	Panel studies to track implementation changes over time Case studies selected to study in-depth project implementation in different contexts Key informant interviews Participant observation Coordination with the project monitoring system
Did people know about the project?	Panel studies Participant observation Social media analysis Building questions into the quantitative impact surveys
Were the right people targeted?	Build questions into impact surveys Compare socio-economic characteristics of project beneficiaries with data from different sources (e.g. previous surveys or census data, academic literature like ethnographies and gender studies) on the characteristics of the total population Participant observation to investigate whether there are any groups such as illegal squatters who live in the project areas but who have not been detected by the project agency Satellite images to identify any geographical groups not included in the project Focus groups and PRAs
Was the treatment correctly applied?	Participant observation Case studies Panel studies Compare project records with observation on the ground
Did exposure lead to uptake?	Participant observation Monitoring data (if available) Key informant interviews Focus groups
Are monitoring data credible and reliable?	Compare monitoring data with direct observation Key informant interviews Focus groups
What to do when surprises occur and unanticipated outcomes are detected?	Try to anticipate unintended outcomes in the theory of change and results chain Use panel studies and case studies to detect unintended outcomes

2.3 Organizational structures and processes

This dimension of the process evaluation assesses the institutional structures of the implementing organizations and the implementation processes. It may include "assessing the internal dynamics of implementing organizations, their policy instruments, their service delivery mechanisms, their management practices, and the linkages among these," as in line with Organisation for Economic Co-operation and Development/Development Assistance Committee's (OECD/DAC's) definition of process evaluation.

The organizational and administrative structures become relevant in a process evaluation to the extent that they have affected the project implementation and results. A comprehensive assessment of organizational processes might not be necessary; the focus should be on the areas of the organization's capacity for project implementation that may affect project results. Sometimes, one will find that the reason for the lack of results lie with the donor structure or the organizational processes of the implementation arrangements. Many evaluations only look at the administrative systems within an agency, but when several agencies are involved, it is equally important to study the coordination and communication between them. There is often a significant level of rivalry between national- and local-level agencies. There may also be coordination issues (and competition) between donors.

In addition, organizational inertia may affect intervention outcomes. Large bureaucracies have established administrative systems, both formal and informal, that they are reluctant to change to accommodate different approaches proposed for a donor project. Inertia can also affect the institutions an agency feels comfortable coordinating with, as well as those with whom there have been difficulties or simply a lack of contact. Sometimes government agencies will reluctantly agree to the new organizational systems and coordination arrangements the funding agency proposes, but the agencies' networks continue informal coordination and communication using their established networks and alliances. Consequently, it is important for the evaluation to address these kinds of institutional inertia in the evaluation.

2.3.1 Examples

Many United Nations (UN) organizations operate with annual project budgets, even in multi-year interventions. This is mainly due to their donor structure. National governments who donate to the UN organizations can often only commit funds for one year at a time, because they are dependent on annual parliamentary approvals. This means that funds are only allocated and disbursed to those organizations on an annual basis and only after national parliamentary approvals. There may be a delay in disbursements from national governments to the organizations and a further delay in disbursements from the organization to the implementing local NGOs (or other implementing partners). This has meant that in some multi-year interventions—e.g. the United Nations Children's Fund's (UNICEF's) education projects in Guatemala, local NGOs have had to close down project activities for up to six months per year while waiting for funds. Evidently, this affects both project implementation and project results.

A large population project in Bangladesh provided an example of challenges of coordination among donor agencies. Bangladesh was a priority country for many donors and all agencies wished to be actively involved. Consequently, on at least one occasion, there was a delay of up to six months trying to find a travel date that would be convenient for all donors for a meeting,[24] and important decisions could not be taken until the meeting had taken place. Also, for the country population project, it was a major challenge to arrange transport for up to 10 donor agency representatives to visit projects and find appropriate meeting locations.

Another example can be drawn from 3ie's process evaluation of the Women's Advancement in Rural Development and Agriculture project (WARDA),[25] where the study team found weaknesses in the institutional structure that affected project delivery and subsequently, the results. WARDA is a technical assistance project helping to establish scalable agri-based value chains that link farmers to markets in India. The project works with SHGs to economically empower smallholder women farmers and increase their agricultural income through market-led interventions.

3ie found that although the donor organization's commitment to building sustainable farmer producer companies had been a crucial enabling factor in reviving them, administrative issues, such as staff turnover and delays in payment to the community cadre, had adversely affected the project. Furthermore, the contractual nature of project staff and the large number of vacant positions limited the project's ability to achieve its objectives. These are all institutional structures and processes that directly or indirectly affect the project's implementation and results.

2.3.2 Coordination and complexity

Coordination among stakeholders is one of the dimensions of complexity included in the complexity map (see **Figure 7**).

A large, multi-level project may involve 100 or more organizational units, many of whom have different and often, inflexible organizational structures, reporting systems and priorities. Some agencies traditionally do not cooperate, while others may be in direct competition for funding. Added to this is the political dimension, where some districts or cities could be controlled by opposition parties that may actively oppose the project. A challenge for evaluators is that while the project may have a formal organizational structure, many organizations actually operate through informal and undocumented structures that are difficult to identify or track.

Systems analysis, such as social network analysis, provides a set of research tools that can be used to address these questions (See Section 3.5). Another potentially useful tool is process tracing (Beach and Pedersen 2013). Process tracing operates at the level of a single case (such as an individual implement ting agency or a single project location) and models the theory a project is based on by articulating the steps through which outcomes are to be achieved. The approach can be applied in three complementary ways: (1) theory testing, (2) theory building and (3) explaining outcome. Data are continually updated to improve the assessment of how closely the theory corresponds to the

[24] This is less relevant now, with the surge of digital meetings during and after the COVID-19 pandemic.

[25] See: https://www.3ieimpact.org/evidence-hub/publications/other-evaluations/evaluating-womens-advancement-rural-development-and

observations on the ground. The approach seeks to articulate and monitor the processes through which the project is implemented, which makes it a useful tool for process evaluation.

2.3.3 Relevant evaluation questions to be considered:
The following list is not intended to be a complete list of relevant questions for process evaluations. They are suggestions evaluators might want to include in their evaluation design and to spark further ideas.

- What was the originally suggested organizational structure of the implementing organizations?
- What deviations (if any) were there from the organizational plan, and what were the reasons?
- Were there any bottlenecks in project delivery? If so, what were they and what were the reasons?
- How is the working relationship between donors and implementing partners, and between different implementing partners (in the case there are more than one)?
- How did the organizational structure affect project implementation and results?
- How is the flow of funds? Is it appropriate to ensure smooth implementation of the project?
- Are relevant policies in place to ensure adequate/appropriate project implementation?
- What about the reporting structure? Do monitoring and reporting capture central/crucial implementation elements that might need attention (e.g. restructuring, additional financing, etc.)?

2.3.4 This may be done by:
Complexity mapping, systems analysis and process tracing are useful tools when assessing the institutional structures and processes with the view of how these may have affected the intervention outcomes, in addition to:

- Reviewing administrative structures/guidelines/regulations/policy documents, etc.
- Interviewing donor representatives
- Interviewing staff and implementing partners in the field
- Reviewing project documentation, budgets, monitoring reports and other correspondence

In *Table 6*, we have inserted the most relevant evaluation questions into a table and linked them to relevant tools for evaluators to consider referencing to where descriptions of the tools may be found.

Table 6: Examples of research tools that can be used to address the different evaluation questions on institutional structures and processes

Question	Tools
Implementation fidelity: What was the originally suggested organizational structure of the implementing organizations? What is the actual organizational structure between donor organization(s) and implementing partners (administrative, economic, etc.), and is it conducive to achieving the objectives? What deviations (if any) were there from the organizational plan? What were the reasons for the deviations?	Complexity mapping Key informant interviews Systems mapping Systems analysis Social network analysis Big data Participant observation
How adequate was the original implementation design (even when correctly implemented) for achieving key project objectives, such as: (1) reaching and serving all sectors of the target population, (2) achieving both social and economic objectives and (3) involving the community in the design and implementation of the project?	Efficiency analysis of the original implementation design Focus groups and key informant interviews Social media analysis
Were there any bottlenecks in project delivery? If so, what were they?	Bottleneck analysis Key informant interviews Documents review (of project documents) Focus groups Etc.
How is the flow of funds? Is it appropriate to ensure smooth implementation of the project?	Key informant interviews Budgets and financial data Flow-of-funds tracking
Are relevant policies in place to ensure adequate/appropriate project implementation?	Analysis of policies, strategies and operational manuals
Reporting structure: Do monitoring and reporting capture central/crucial implementation elements that might need attention (e.g. restructuring, additional financing, etc.)?	Monitoring data Management Information System Analysis of theory of change to identify the most relevant indicators Key informant interviews Focus groups Observation: How is monitoring information actually collected and recorded?

2.4 Context and external factors

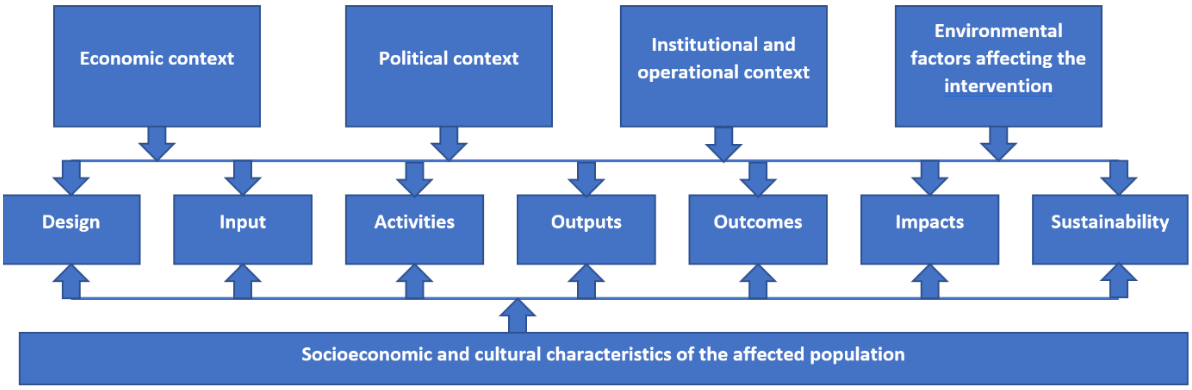
Development interventions never operate in a vacuum and are always subject to external factors, such as the local, regional and wider economic, political, institutional and environmental contexts.[26] In addition, the specific socioeconomic and cultural traits of the affected communities may be a determining influence for the success of the intervention,

[26] See also Hentschel (1999), Patton (2002), Bamberger and Mabry (2020), and Gertler et al. (2011).

especially if they are not taken into account during planning and implementation. If the evaluation disregards external factors and socioeconomic and cultural characteristics, it may miss important causal information and draw the wrong conclusions. Despite the fact that policymakers, funders and development agencies all recognize that project implementation (and project outcomes) are significantly affected by these and other external factors, most process and impact evaluations do not include a framework for the systematic analysis of these factors.

Figure 6 outlines a simple results chain including external factors.

Figure 6: Simple results chain including external factors[27]


Economic context
Political context
Institutional and operational context
Environmental factors affecting the intervention
Design
Input
Activities
Outputs
Outcomes
Impacts
Sustainability
Socioeconomic and cultural characteristics of the affected population

Comprehending the context is crucial to understanding project impact and designing the evaluation. Context means the social, political, economic, and socio-cultural setting the project takes place in; all these can influence how the causal chain plays out. The impact of an identical project can differ in various geographical contexts. Understanding context means a thorough reading of project documents prior to embarking on evaluation design, but also exposure to a broader literature, such as anthropology and political economy. There are also many socio-cultural and political factors that are not documented in publications, so where possible, the initial diagnosis should also include techniques such as key informant interviews, focus groups and observations through project visits. New big data sources such as social media analysis, radio call-in projects and satellite images may also be available.

The following are examples of how each of these contextual variables can affect a development intervention and how their analysis can strengthen the interpretation of evaluation findings:

2.4.1 Economic factors

Economic factors, such a country's economy and its economic system, and the poverty levels of end users of a development intervention could affect project outcomes. For example, in a dynamic economy, where demand for products and services is growing and new jobs are being created, people are often more willing to invest time or resources in developing marketable skills or launching businesses. Parents may also be more

[27] This model is based on Figure 2.2. on page 40 in Bamberger et al. (2006) and Figure 2.2. on page 27 of Bamberger and Mabry (2020), with some modifications for the purpose of these guidelines.

willing to pay for their daughters to stay in school if cultural constraints and labor market conditions create the expectation that education will help the girls be employed and get an income. The degree of poverty could also be a driving factor keeping children out of school, because their families might rely on their children's income from petty work. Similarly, nutrition projects that provide supplementary meals for children might experience that poor families use the food as substitution of a meal that would, otherwise, be given at home rather than as a supplement (i.e. in addition to the home meal). All this will affect project performance. These are important economic (external) conditions that would need to be taken into account both in the design and evaluation of an intervention.

2.4.2 Political factors

Different political factors may also affect project implementation. For example, if the project would promote practices not supported by national policies—e.g. privatization of water companies, it will be difficult to implement unless national policies are revised. Similarly, support from local government agencies that happen to be from the same political party as the national or state government sponsoring a project may significantly improve project performance by mobilizing community support or providing free resources, such as transport, workers or buildings. Inversely, politically induced opposition to a project could seriously affect its success or even its ability to operate. Sometimes, projects can become affected by political campaigns, and changes in political administration (centrally or locally) may also affect project implementation, because new administrations often do not want to "inherit" projects from the old administration.

Government commitment was a key ingredient in the success of a World Bank-financed project to reduce maternal mortality and fertility and improve child health in Bangladesh (World Bank 2005). The country went from almost no facilities immediately after independence to having a nation-wide decentralized health and family planning system, down to doorstep delivery of contraceptive services, in a 10-year period. Similarly ambitious projects may falter if government does not have the will to see them through.

A World Bank self-housing project that one of the authors worked on in Zambia in the 1980s, was designed to include user charges to ensure cost recovery. Cost recovery for water and housing was a new concept for low-cost housing in Zambia, and the ministry only agreed to it after extended negotiations. However, the project planners did not take into consideration the upcoming municipal elections, where the platform of one of the candidates was to provide free housing to combat capitalist exploitation. The promise of free housing discouraged families from paying user charges. The lesson from this example is that had the evaluation taken relevant political factors into account, the team would have probably identified the municipal elections as a factor to be monitored.

Another important, but difficult to measure, set of factors relates to *elite capture*, where benefits targeted for low-income communities may be captured by groups with political contacts even prior to project implementation. A common example is where political insiders buy up land whose value is expected to rise as a result of road construction, slum upgrading or other public infrastructure projects. Often, the evaluation does not capture these clandestine transfers, because they occur before the evaluation begins and the baseline study is conducted.

2.4.3 Organizational and institutional factors
Many projects require support from government agencies and other organizations, such as NGOs, community-based organizations or the private sector. The effectiveness of such inter-agency cooperation can vary considerably from one community or district to another. In some cases, this is due to personalities, in others, to local politics but also differences in staff, financial or other resources. At times, something as basic as the fact that the ministry in one town has a jeep, whereas in the next town, it does not might considerably affect the level of institutional support.

Funding agencies often seek to introduce new coordination mechanisms among agencies that have not previously worked together, or where the new arrangements would be difficult to implement and require new administrative procedures. However, agencies may be reluctant to make these changes and may informally continue their previous practices, while officially following the new procedures. Even if there is no active opposition to the new procedures, inertia and slowness to introduce changes are common. The existence of these informal coordination mechanisms will rarely be documented, so qualitative methods will often be required to understand how coordination among agencies actually works (or in some cases, does not work).

2.4.4 Environmental factors
Agricultural and rural development projects are directly affected by variations in the local environment. A new grain variety may be dependent on chemical, commercial fertilization and not respond to traditional, natural fertilization; it may prosper well on flat land, but not on hillsides; or it may be very sensitive to variations in seasonal rainfall. Similarly, urban development projects could be affected by erosion or flooding, and similar. All these factors may produce dramatic differences in crop yields or the success of water and sanitation projects, for example.

There are also situations where the new systems may be environmentally beneficial for the target population, but create negative environmental consequences to groups not directly involved in the project. One example is where miracle rice in countries such as Indonesia is able to introduce a second or third growing cycle, which significantly increases efficiency for large farmers who can afford the new fertilizers and other investments, but may mean that there is no longer a fallow period when poor farmers can feed their goats at no cost. Similarly, sustainable rural energy projects may mean agricultural waste is used to generate power, but is no longer available to small farmers as fodder for goats and cows. Many cultures have traditional ways to provide the poorest households with access to grazing or fallen fruit, but these survival mechanisms are not documented and are frequently, not taken into consideration when assessing the economic benefits for land privatization and other modern farming practices.

2.4.5 Socioeconomic and cultural characteristics of the target communities
Farming practices, rules concerning use of natural resources, marriage practices, gender and power relations, and attitudes concerning the mobility and economic participation of women vary greatly, not only between countries but often also between areas within a country and between different ethnic groups. In one village in Uganda, bicycles proved an effective way to transport water and reduce women's time burden, because water was carried in square metal jerry cans that could easily be transported on a luggage rack. However, in a neighboring village, bicycles failed to produce this benefit, because

water was transported in round clay pots that could not easily be transported on a luggage rack. A goat scheme in Zambia was meant to auto-target women, but failed to do so because goats were taboo for women living in the project location. The lack of consideration of the power relations between mothers-in-law and daughters-in-law in Bangladesh (as pointed out in the example under section 2.2, *page 19*) led to weaker results in the Bangladeshi nutrition project, as did the lack of attention to the power relations between men and women in the kitchen garden project in Zambia.

An analysis of such socioeconomic and cultural factors can often help explain why two identical projects may have very disparate outcomes in different communities. Such contextual factors are crucial to consider, not only when designing a development intervention, but also during evaluation. As part of a process evaluation, it is important to assess the appropriateness of the problem analysis (diagnosis) and the project design (treatment), in light of context-related factors.

2.4.6 Examples

In southern parts of Zambia, people are cattle-keepers and tend to measure their wealth in number of cattle owned. In those areas, cattle are not only used as assets and an important food source, but also as draft animals to prepare the land for cultivation, which is much more efficient than preparing the land by hand. In the 1990s, Southern Zambia suffered from a severe attack of corridor disease[28] and most of the cattle died, spurring the population into a downward poverty spiral, because they not only lacked meat but also struggled to prepare sufficiently large areas for cultivation by hand, having lost their draft animals; many also found themselves forced to sell off farm assets, such as ploughs, to get money for food. An agricultural development organization that had previously supported the area with cattle projects found themselves unable to continue working in the area, because cattle could not be reintroduced again until the end of a 10-year quarantine period. Instead, they decided to introduce cattle as draft animals to the people living in the forest areas of Northern Zambia—a population who until then, had never owned cattle and had only prepared their land by hand. The project's critical assumption was that land preparation by draft animals is much more efficient and that introducing cattle in Northern Zambia would, thus, lead to increased food production and improved food security in the region. The project looked good on paper; all animals had been distributed and farmers seemed to be happy. When one of the authors arrived there with a study team a couple of years into project implementation, we found that most people feared the big animals and did not know how to treat and use them. The animals were kept in fenced off areas, away from people, and people threw food to them without getting too close. Very few (if any) of the animals were used to prepare the land. It was also unclear to what extent the animals were used for food consumption, because this was not familiar food in the area.

This is a typical example of a project where cultural and environmental contexts have not been taken into account in project design. The project worked very well in Southern Zambia until they were hit by the corridor disease, but did not work in an area where people lived in the forest and were not cattle keepers.

[28] Corridor disease is an acute, usually fatal disease of cattle, resembling East Coast fever. It is caused by infection with buffalo-derived Theileria parva strains transmitted by ticks from African buffaloes.

In his book “Guinea-pigs: food, symbol and conflict of knowledge in Ecuador” (1997), the anthropologist Eduardo Archetti explains why a World Bank food security project with the aim of scaling up the production of Guinea pigs in the rural areas of Ecuador failed to improve people’s nutritional status. Indigenous people in the Andes have reared and eaten Guinea pigs since ancient times, so it seemed only rational to development planners to “modernize” their production for increased consumption. The underlying assumption was that increased production of Guinea pigs would lead to increased consumption and improved food security. When the intervention failed, Archetti and his colleagues were recruited to look at why it failed. They found that Guinea pigs carry a meaning in the social and ritual life of Ecuadorian peasants, which is far from mundane. They are consumed only during religious rituals and festivals, and people who keep them do not categorize them as food.

2.4.7 Heterogeneity and generalization

In the present context, heterogeneity refers to the fact that projects are often implemented in many different communities or geographical regions, each with unique characteristics such as the ethnic and demographic characteristics, local economy, availability of different kinds of infrastructure and public services, and local political situation, among others. All these can affect project implementation and outcomes. Consequently, a project that follows the same implementation plan can have very different results depending on the influence of these multiple factors.

Heterogeneity can also be important in assessing the potential project replicability in other contexts. For example, what are the characteristics of households, communities, organizations, and so on in relation to project success. New locations can then be assessed in terms of the proportion of households with these characteristics. It should be noted that randomized controlled trials (RCTs) and other regression-based research methodologies are usually designed to estimate the likely impact of a project after controlling for these factors. In other words, an RCT evaluation is useful for assessing the impact of a project in a particular context, but limits the ability to provide guidance on potential replicability—how the project would be likely to perform in other contexts with a different configuration of the socio-cultural, economic and other factors. The message of these guidelines is that rigorous quantitative tools, such as RCTs, are a valuable component of the project evaluation process, but for many operational planning purposes (such as replicability), they should be part of a mixed-methods approach that incorporates other research methodologies, both quantitative and qualitative. See review of evaluation methodologies in Section 3, *page 37*.

While a heterogeneity analysis would be undertaken as part of an impact evaluation, understanding the context may help explain possible impact heterogeneity. Impact (i.e. the treatment effect) can vary according to intervention design, beneficiary characteristic or the socioeconomic setting. Examining the underlying theory can help expose possible heterogeneity and allow the evaluation design to anticipate it.[29]

As outlined by White (World Bank 2005), in child feeding projects, for example, malnourished children are more likely to respond with weight gains than well-nourished children, although extremely poorly nourished children may have diarrhea, which

[29] See also White and Vajja (2008) and World Bank (2005) for a further analysis of this.

prevents effective feeding and weight gain. Better targeted projects will, thus, have a higher average impact, and that impact will be the greatest in the lean season. Younger children are likely to benefit more than older children, because children who have suffered stunted growth in infancy will not experience marked height gains from feeding in later years. Similarly, cognitive gains from better nutrition appear to be captured under three years of age. Hence, impact varies by beneficiary age and preexisting nutritional status, the latter having a seasonal element. Impact can also vary according to socioeconomic status—e.g. substitution (using supplementary feeding to replace an existing meal) is more likely in poor households.

Another aspect of heterogeneity is the possible complementarity between interventions—e.g. as pointed out by White and Vajja (2008), it may be that microfinance has a large impact if accompanied by business support services. Or maybe the two are substitutes, where the impact of the two combined is less than the sum of the two separately. Designs that explore such complementarities can be of great policy relevance.

Understanding context can also help generalization. For example, studies of World Bank's support to basic education in Ghana and of maternal and child health in Bangladesh were overall success stories. In the Ghanaian case, large-scale school rehabilitation and textbook provision made significant contributions to improved enrolments and learning outcomes (World Bank 2004). There were two important contextual aspects behind this result. First, following years of crisis, the school system was in a truly poor state, with inadequate infrastructure and virtually, no school supplies. School renovation and textbook supply had an impact in this context that it may not have had if schools had already been relatively well functioning. Second, there was a strong political support for the project, which helped ease implementation, because the project was part of a wider educational reform.

The main questions to be considered under context, external factors and heterogeneity are:

- What are the contextual factors that might have affected the design, implementation and outcomes of the development intervention?
- What might cause issues with heterogeneity?

Useful tools to address those questions are a literature review of relevant academic publications, such as anthropological ethnographies, gender studies, economic and environmental studies, and previous evaluations from the project area. Key informant interviews may also be useful, as would direct and participant observation in the field, including focus groups and various PRA techniques.

3. Evaluation designs for integrated process evaluations in impact evaluations

While this report focuses on the use of process evaluation to strengthen impact evaluations, process evaluations are also useful for a number of other purposes. For example, process evaluation can also provide feedback to management to strengthen ongoing projects, help ensure that no groups are being excluded or improve the design of future projects.

However, there is no one recipe for how to conduct a process evaluation or incorporate relevant process evaluation elements into impact evaluation designs. Each process evaluation needs to be tailor-made to the object of study, taking into account the context, sector, problem analysis and project design, purpose of the evaluation, and the research problem and questions.

Before delving into the three main scenarios under which process evaluations are conducted to strengthen impact evaluations (from sections 3.3 onwards), we would like to outline a couple of important points. Firstly, the evaluation should address the information needs of key stakeholders, and the design should be problem-driven—your information need (what is it that you need to know) is what drives the identification of the most appropriate methods to provide you that information. Secondly, we recommend the application of a mixed-methods framework (see Section 3.2, *below*), because it permits the evaluator to draw on the widest possible range of evaluation methods and tools, increase the validity of conclusions, and provide a deeper and richer analysis and interpretation of the context.

3.1 Match the design to the research problem, purpose and questions

Prior to the start of any evaluation, the first thing to do is to clarify the purpose of the evaluation and understand the information needs of different stakeholders. A key issue is to define which groups will be consulted: Is it only the client? Only donors and a few government agencies? Will the intended beneficiaries and the affected populations also be consulted? It is surprising how many evaluations pay little attention to the affected populations in the design, implementation and interpretation of the evaluation. There are well-established procedures for stakeholder analysis that can be consulted (Greene 1997, 2005; Patton 2008).

Once the issue of stakeholders has been clarified, the next question concerns what we *need to know*, which is different from what it would be *interesting to know*. What are the main questions we need answered? Evaluation methods may then be selected based on what kind of information each method can provide answers to.[30] *Table 7* shows a data matrix that can be used to translate key information needs into a set of questions to be included in the data collection instruments. For each question, various possible data sources are listed, and each is assessed in terms of the feasibility of collecting the data and their reliability. Column 4 identifies the preferred data collection method and Column 5 provides a backup option.

As White (2009:61) points out, evaluations should be issues-driven, not methods-driven. Many evaluators and many clients have a preference for certain evaluation methodologies, and some terms of reference will specify the preferred, or even required, methodologies.[31] One of the advantages of a mixed-methods approach is that the

[30] While this is commonly accepted research practice for most researchers, this is not always the case in evaluations. Often, the evaluators are responding to terms of reference written by operations staff without a research background, where a “standard” methodology is prescribed: documents review, interviews, and focal group discussions, for example.

[31] Barbrook-Johnson et al. (2021) report that the United Kingdom government evaluation procurement procedures present a major barrier to the incorporation of complexity-responsive evaluation methods, because the research proposal is required to define precisely all stages of the

balance between different methods can be adjusted according to the specific information needs of the study and the stakeholder's concerns. This means that the starting point should be the *evaluand*—the project to be evaluated or the policy to be tested, and the stakeholders' information concerns.

Table 7: Section of a data matrix for a process evaluation

Evaluation question	Possible Indicators	Possible data sources	Adequacy and accessibility of each option	Alternative (less reliable) data sources	Comment/ decision
1. How well does the project design ensure that all sectors of the community will benefit from the project?	a. Sex, ethnicity and other characteristics of people involved in project meetings and workdays?	a. Attendance records for community meetings	a. Not reliable, because it is often completed by memory, after the meeting	a. Focus groups b. Key informant interviews	(1) Use 3 data sources for a 3-month period and organize a team meeting to assess reliability and feasibility of each source (2) Compare results with focus groups and/or key informants
		b. Attendance records for workdays	b. Quite accurate, but only indicates sex		
		c. Observer reports on meetings and workdays	c. Observers required to complete checklist covering most of the attributes, but estimates of ethnicity are not very reliable		
2. How actively did different groups participate and how actively were they involved in decisions and management?	a. Membership of different groups in committees and other positions	a. Meeting attendance list b. Group membership list c. Observing meetings with a key informant who can identify the social groups each participant belongs to	a. Most attendance lists do not report social groups attendees belong to b. More reliable, but time consuming		
	b. Contributions of each group (food, money,	a. Minutes of meetings b. Records	a. Do not always report contributions b. May only record		

evaluation, which makes it impossible to incorporate the flexibility required to address the complex and changing environments where programs and policies are implemented.

Evaluation question	Possible Indicators	Possible data sources	Adequacy and accessibility of each option	Alternative (less reliable) data sources	Comment/ decision
	lending tools etc.)	of contributions	some kinds of contributions (e.g. money and labor), but not others (e.g. food contributions or time spent on coordination)		
		c. Observing meetings	More reliable, but time consuming		

3.2 Three scenarios where process evaluations are integrated into impact evaluations

There are three main scenarios under which process evaluations can strengthen impact evaluations (see *Table 4*). This section describes each scenario, when they are used and their main purpose. The table also includes examples of evaluation designs for each scenario. Later, in Section 3.6, there is a more extensive discussion of 14 different evaluation designs, and their potential applicability to each scenario (see *Table 7*).

3.2.1 Scenario 1: Retrospective integrated process/impact evaluation designs

These evaluations are conducted after a project is implemented. They may be conducted as part of the end-of-project reporting or one to two years after project completion.[32] A limitation of retrospective evaluations is that it is not possible to directly observe the implementation process, so the analysis must rely on reports produced by the implementing agencies and possibly, partner agencies, and ex-post interviews with project agencies, target populations and key informants. While documentation is almost always available on the *formal* implementation process, it is difficult to obtain reliable and representative information on what actually happened during implementation, which can be significantly different from the formal project implementation plan. There is a danger of obtaining a positive bias, because most agencies are reluctant to report on things that went wrong and may not even know whether there were any groups (such as ethnic minorities, illegal squatters, refugees) who were excluded (intentionally or unintentionally) from access to project benefits. The present authors have been involved in project evaluations where the project communities included more than 25 percent of illegal squatters whose presence was unknown to project management. In some settlements, illegal squatters would hide whenever agency staff or other outsiders visited the project (see Salmen 1987 for an example from Bolivia).

There are a range of data collection methodologies (see Section 3.6 and Annex 2) to address the problems of positive bias and under-reporting, but many of the techniques are time-consuming and require a higher level of interview skills than what is needed for the administration of structured survey instruments. A central message of this guideline is that for most impact evaluations, it is essential to use a mixed-methods design that can

[32] Many agencies conduct a routine end-of-program evaluation for all programs and then, select a sub-sample one or two years after completion for a more intensive retrospective impact evaluation.

combine a range of quantitative and qualitative methods and uses triangulation, comparing estimates from different data sources to strengthen validity and control for bias.

3.2.2 Scenario 2: Pretest–posttest designs

These designs compare baseline and end-of-project data to assess the changes that took place, and how the process and quality of implementation affected intended outcomes. Most quantitative impact evaluations use either RCTs or quasi-experimental designs with a statistically matched control group. Majority of quasi-experimental designs use either naturally occurring experiments (e.g. where projects are delayed in some areas due to floods or administrative delays) or planned variations, as when a new project or service is introduced in phases in different regions. Often, sampling techniques such as propensity-score matching or mechanism experiments, are used to avoid sample selection bias. Due to budget or time constraints, some quasi-experimental designs are forced to use judgmental sample selection to match the project and comparison groups.

In many of these designs, time and resources do not permit the collection of primary data on the process of project implementation and often, the process evaluation component is based on quantitative data collected during the ex-post surveys. However, there are many aspects of project implementation that cannot be understood solely through quantitative surveys and, when resources permit, there are important benefits from the collection of primary data through techniques such as observation, in-depth interviews, focus groups and participatory group consultations.

All of the evaluation designs can be significantly strengthened if they can coordinate with project management to integrate data from the project monitoring system into the evaluation.

3.2.3 Scenario 3: Formative/real-time process evaluation designs

Recognizing that project implementation is a complex and dynamic process, with important differences between the implementation protocol (project design) and what actually happens on the ground, the formative/real-time evaluation approach uses a range of data collection and analysis techniques to directly observe how the project is actually implemented. This builds on the well-established formative evaluation approach (Scriven 1991, Rossi, Lipsey and Freeman 2004, Patton 2008). In addition to strengthening the end-of-project assessment of how implementation affects outcomes, formative evaluation also provides ongoing feedback to project management to identify and correct implementation problems in real-time.

Formative/real-time evaluations use an eclectic approach, combining many different sources of available information that reflect the specific characteristics of the project and its environment. As for the other scenarios, it is recommended that this design be implemented within a mixed-methods approach.

There are several features that distinguish formative/real-time process evaluation from the two earlier scenarios.

- While the two previous approaches tend to focus on the formal project design and how it is implemented, formative/real-time evaluation has a broader focus on the socio-cultural and political environment the project is implemented in, as well as the broader economic, political, administrative, demographic and climatic environment, and how they affect project implementation.

- The focus is on the *process* of project implementation (using observation and other continuous measurement techniques), rather than measurements taken at only a few points in time.
- There is a focus on the informal socio-cultural and political factors that determine how implementation actually takes place, rather than a narrow focus on intended project design.

The analysis addresses issues of social exclusion and seeks to identify groups that are excluded from, or only have limited access to, the project due to issues of race, ethnicity, gender, refugee status or political affiliation.

Table 8: Integrated process/impact evaluation scenarios

	Scenario		
	1. Retrospective	**2. Pretest–Posttest**	**3. Formative/Real-time**
When is this conducted?	End of project or 1 to 2 years after project completion	Start and end of project, possibly with a mid-term measurement	Ongoing, throughout project; In some cases, initial diagnostic study can also cover a period before the project officially begins
Purpose	• Assessing fidelity of project implementation • Assessing how fidelity of implementation affected outcomes • Identifying any excluded or under-served groups	• Strengthening assessment of effects of implementation and variations in implementation on project outcomes • Identifying any excluded or under-served groups • Assessing how implementation and outcomes are affected by the context the project operates in	• Comparing intended implementation strategy with what actually happened on the ground • Providing real-time feedback to managers to strengthen ongoing implementation • Identifying any excluded or under-served groups • Assessing how implementation and outcomes are affected by the context the project operates in
Recommended evaluation designs	There are 3 design recommendations that should be considered and where possible, applied in all evaluation designs: a. *All evaluations should incorporate a mixed-methods design that uses triangulation to obtain and reconcile at least two independent estimated key variables. Mixed methods also incorporates qualitative methods to strengthen the statistical analysis.* b. *All evaluations should incorporate a complexity analysis, which can vary from a simple complexity map to more refined applications of systems analysis techniques, such as systems mapping, system dynamics, social network analysis or boundary analysis (critical systems heuristics)* c. *All evaluations should consider the feasibility and utility of incorporating big data sources and analytical techniques (see Section 3.7.4)*		

	Scenario		
	1. Retrospective	**2. Pretest–Posttest**	**3. Formative/Real-time**
	• Ex-post experimental designs, combined with fidelity analysis or implementation research to obtain quantitative assessments of project implementation; The analysis estimates the influence of implementation variations on project outcomes • Quasi-experimental designs and natural experiments (using propensity–score matching when experimental designs are not possible • Focus groups, key informant interviews and project visits may also be incorporated • Direct observation and participant observation, PRAs, etc.	• Pretest–posttest comparison group designs, sometimes incorporating a mid-term measurement • Randomized designs are used where possible, but often only a quasi-experimental design is possible • Where possible, mechanism experiments are incorporated to manipulate components of the implementation design and assess effects on project outcomes • Where possible the statistical design is complemented by qualitative techniques to test for variations in how the project is actually implemented • Focus groups, direct observation, participant observations, PRAs, etc. (here, the *rule of thumb* is relevant)	• This scenario can incorporate all the techniques used in Scenario 2, but with an additional focus on providing feedback to project management on how to detect and address in real-time problems affecting project implementation • This formative objective often involves more time in the field to observe what actually happens during implementation, as well as working with management to incorporate monitoring and other kinds of management information systems into the evaluation design • There is also greater attention to processes of emergence and how the implementation adapts to changes in the organizational, political, economic and socio-cultural environment within which the project operates

3.2.4 Refining the integrated process/impact evaluation scenarios

The above scenarios provide a useful framework for initiating process evaluations as part of impact evaluations. However, there are two refinements that can be included for a more in-depth and focused process. The design and analysis of the process evaluation should be adapted to both of these factors.

First, developing a typology of intervention levels and identifying the required refinements to the process evaluation design for each type. The typology might include: (1) small localized projects, (2) larger projects with different components and a wider geographical coverage, (3) sector-wide projects (such as educational or health reforms), (4) country-level projects, (5) multi-country and global projects (such as migration control or climate change) and (6) policy interventions. Second, adapting the process evaluation to the unique characteristics of different development sectors. Implementation strategies vary significantly by sector.

3.2.5 Integrating the findings of the process evaluation into the impact evaluation design

In all three scenarios, most of the process evaluation data are qualitative. There are two main ways the findings of the process evaluation can be used to strengthen the impact evaluation. The first is to use the rich descriptive data to help explain some of the variations in the level of change in the outcome indicators. The second is to refine the analysis by transforming the qualitative data into a set of ordinal rating scales that can strengthen the rigor of the assessment of the influence of the quality of implementation on the outcome variables.

3.2.6 The key features of the integrated process/impact evaluation approach (Figure 6)

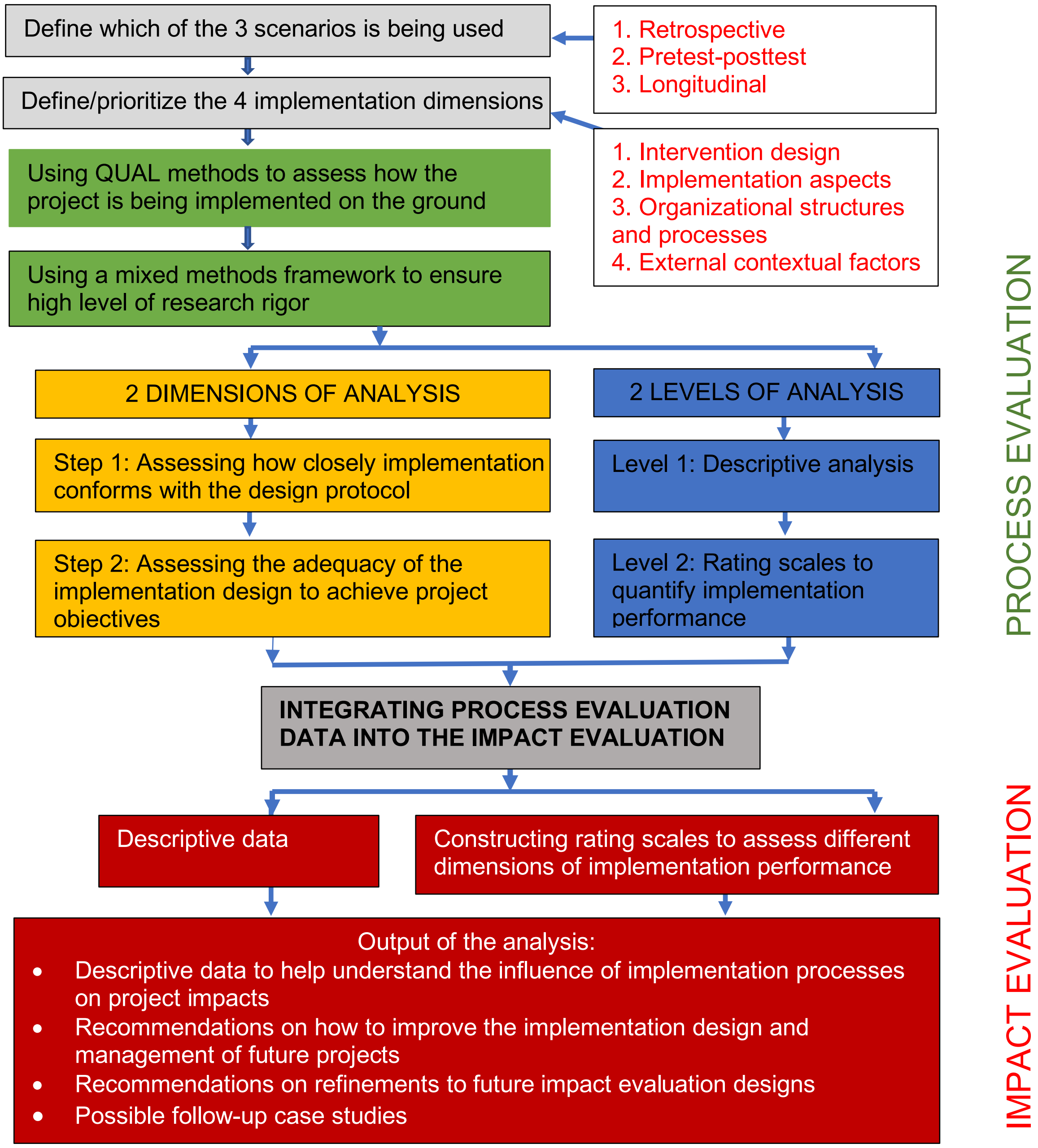
While each of the three scenarios described in the previous section have certain unique features, all the designs have the same underlying logic:

1. The purpose of all of the process evaluation designs is to strengthen the impact evaluation by incorporating information on how the project was implemented and using this to help interpret how the quality of implementation affects project outcomes and impacts.
2. Project implementation is an ongoing dynamic process that is influenced by a wide range of political, organizational, socio-cultural and external contextual factors. What actually happens on the ground during implementation often differs significantly from the official implementation plan. In most projects, the information on these real-world processes is not adequately documented in project monitoring systems and other project reports, and can only be fully understood through a creative combination of quantitative and qualitative collection and analysis methods. The wide range of potential research tools that can be drawn upon are summarized in Annex 2.
3. When combining research tools with different research frameworks that use different approaches to quality control, it is important to maintain a high level of methodological rigor that reconciles the different approaches. The guidelines recommend basing the evaluation on a mixed-methods evaluation framework. Mixed-methods approach systematically combines appropriate quantitative and qualitative techniques at each stage of the evaluation process, and considers ways to reconcile different philosophical and methodological approaches. There is a strong emphasis on triangulation and other forms of quality control to ensure construct validity and reliability. However, there are important methodological and analytical issues that require further discussion.
4. There are two main dimensions of process evaluation that are assessed. First, how well did project implementation comply with the implementation design (implementation fidelity). Second, the adequacy of the project implementation design for achieving all project objectives. The distinction is important, because there are situations where project implementation closely follows the implementation protocol, but this may not be adequate to achieve all project goals. There may be some goals that are not addressed or not addressed

adequately. For example, the implementation design may work well for better-off families, but not so well for poor families, female-headed households or farmers who do not own their land. The issue of emergence can be important, because the implementation might work at the start of the project, but not have the capacity to adapt to changing circumstances.

5. The main way process evaluation can strengthen impact evaluation is by providing detailed descriptive information to help understand how the different ways a project is implemented can affect outcomes and impacts. This information can aid in explaining variations in project impacts and provide illustrations through case studies, observation and in-depth interviews. It can also help identify areas for further research or modification of the design of future impact evaluations.
6. It is also possible to summarize the assessment of the quality of project implementation in a set of rating scales. For example, a set of indicators can be defined to assess implementation. These might include: (1) How closely did implementation follow the project's implementation protocol?, (2) How adequately did the implementation plan cover all the project's goals?, (3) How well did implementation ensure the inclusion of all sectors of the target population?, and (4) How efficient and cost-effective was implementation? Performance on each of these dimensions can be rated on a set of scales, using ratings such as: 5=highly satisfactory, 4=satisfactory, 3=adequate, 2=poor, 1=very poor. Where appropriate, the scales can be incorporated into the impact evaluation reports in the ways that OECD/DAC evaluation rating scales are widely used.
7. It is recognized that more work is required to explore ways to strengthen the rigor and utilization of the wide and rich range of process evaluation tools in impact evaluations.

Figure 7: The integrated process/impact evaluation design


Define which of the 3 scenarios is being used
1. Retrospective
2. Pretest-posttest
3. Longitudinal
Define/prioritize the 4 implementation dimensions
1. Intervention design
2. Implementation aspects
3. Organizational structures and processes
4. External contextual factors
Using QUAL methods to assess how the project is being implemented on the ground
Using a mixed methods framework to ensure high level of research rigor
2 DIMENSIONS OF ANALYSIS
Step 1: Assessing how closely implementation conforms with the design protocol
Step 2: Assessing the adequacy of the implementation design to achieve project objectives
2 LEVELS OF ANALYSIS
Level 1: Descriptive analysis
Level 2: Rating scales to quantify implementation performance
PROCESS EVALUATION
INTEGRATING PROCESS EVALUATION DATA INTO THE IMPACT EVALUATION
Descriptive data
Constructing rating scales to assess different dimensions of implementation performance
Output of the analysis:
Descriptive data to help understand the influence of implementation processes on project impacts
Recommendations on how to improve the implementation design and management of future projects
Recommendations on refinements to future impact evaluation designs
Possible follow-up case studies
IMPACT EVALUATION

3.3 Key questions addressed in each evaluation scenario

As discussed in Sections 2.1, 2.2 and 2.3 *above*, there are a number of key questions that are addressed in all process evaluations. They include:

1. How adequate was the project diagnosis, and how well did the project design address the major problems and priorities of the target population?
2. How closely did implementation comply with project design?
3. How did the level of adherence to project design affect outcomes in general and for different sectors of the target population?
4. How did the institutional structures and processes affect project outcomes?
5. What are the contextual factors that affect implementation and outcomes?

Table 9 gives examples of additional questions that can be included in each of the three evaluation scenarios.

Table 9: Key questions addressed in all process evaluation scenarios and specific questions for each scenario

Key questions included in all process evaluations	**Additional questions often included in the different scenarios**		
	1. Retrospective/ Ex-post evaluation	**2. Pretest–posttest comparison evaluation**	**3. Formative/continuous evaluation**
1. How adequate was the project diagnosis, and how well did the project design address the major problems and priorities of the target population? 2. How closely did implementation comply with project design? 3. How did the level of adherence to project design affect outcomes in general and for different sectors of the target population? 4. What are the contextual factors that affect implementation and outcomes? In addition, each scenario usually focuses on a set of scenario-specific questions (see following columns)	1. How did refinements to project design components affect outcomes? 2. How did the dimensions of complexity affect project implementation and outcomes?	1. How did organizational and administrative arrangements affect project implementation and outcomes? 2. How did refinements to project design components affect outcomes? 3. What was the project rating on each of the four dimensions of complexity (defined in the complexity map), and how did the complexity ratings affect implementation and outcomes?	1. How closely does/did the actual implementation process correspond to project design? 2. What are/were the factors causing implementation to deviate from the project design? 3. How does/did the actual implementation process affect the inclusion or exclusion of different vulnerable groups? 4. How did socio-cultural factors affect the actual process of implementation? 5. What were the main dimensions of emergence and how adequately did the implementation strategy identify and adjust to these trends?

3.4 Designs for integrated process/impact evaluations and their applicability in each evaluation scenario

Section 3.3 identified the three main scenarios for conducting integrated process/impact evaluations.

1. Process evaluation combined with a retrospective impact evaluation: Under this scenario, both the process and impact evaluations are conducted at the end of the project.
2. Process evaluation combined with a pretest–posttest impact evaluation that uses either a randomized or quasi-experimental design: With this option, the process

evaluation can either be conducted longitudinally throughout the project or cover a more limited time period. The process evaluation can be mainly quantitative, drawing on secondary data sources or primary survey data, or combine quantitative and qualitative methods. In some cases, it is also possible to use mechanism experiments, where randomized designs can test variations on the components of the implementation strategy.

3. Pretest–posttest experimental design can be combined with a formative process evaluation that provides feedback to management to improve ongoing implementation and understand how implementation affects project outcomes.

Each of the above designs normally addresses all of the following questions, although the importance given to each question can vary.

1. The appropriateness of the diagnosis and the prescribed treatment
2. The efficiency and effectiveness of project implementation, factors affecting implementation, and how the different dimensions of implementation effectiveness affect project outcomes and impacts
3. The influence of institutional processes and structures on implementation and outcomes
4. Contextual factors affecting implementation
5. The efficiency of management information systems to process feedback from the evaluations and adjust implementation systems in response to the feedback

3.4.1 The importance of mixed methods

Table 11 describes 14 integrated process/impact evaluation designs. It is recommended that each of these designs be implemented within a mixed-methods evaluation framework that incorporates both quantitative and qualitative data collection and analysis tools. A mixed-methods approach combines the statistical rigor and capacity to generalize findings to the total population (controlling for selection bias) of quantitative methods, with the ability of qualitative methods to observe ongoing processes, provide in-depth description of individuals and groups, and understand the opinions and values of different sub-groups. Mixed-methods approach also helps understand informal power relations and other contextual factors that may affect project outcomes. It combines counterfactual analysis (the underlying logic of experimental designs) with analysis of the factual (the logic of the project design and how it is actually implemented).

An additional benefit of a mixed-methods approach is the use of *triangulation* (see Section 3.7.3), which compares and reconciles estimates of key information from two or more independent sources, thereby strengthening validity and increasing the credibility of findings to different audiences, some of whom may have a preference for quantitative data, while others have more confidence in qualitative data.

It is important to note that many quantitative impact evaluations already incorporate some qualitative methods (e.g. focus groups, key informant interviews, project site visits and case studies). However, the qualitative approaches are frequently used in a somewhat ad hoc way, due to budget and time constraints, and triangulation is usually not incorporated systematically. The recommended mixed-methods frameworks is to systematize current practice and not introduce a radically new approach.

Mixed-methods designs are flexible and must always be adapted to the characteristics of each implementation process to take advantage of the different kinds of data available.

3.4.2 Design options for integrated implementation/impact evaluation designs

The recommended designs are the following (see *Table 7*):

1. ***Mixed-methods evaluation framework.*** This framework combines quantitative generalizability and statistical rigor with qualitative depth of analysis of behavioral change and tracking of processes of change. Mixed-methods designs also use triangulation to independently compare at least two separate estimates of key variables to strengthen validity and credibility. Mixed-methods approaches were discussed earlier in this section.
2. ***Mapping complexity and assessing the influence of contextual factors and organizational arrangements on implementation and outcomes***. The complexity map (*Figure 8*) identifies four dimensions of complexity that individually, or in combination, influence project implementation and the achievement of outcomes and impacts.[33] It is recommended that a simple complexity map be part of the design framework for all process evaluations to help identify a wide range of factors that can potentially influence implementation and outcomes. The map serves as a checklist to remind evaluators of important influences that are largely ignored in many evaluations.

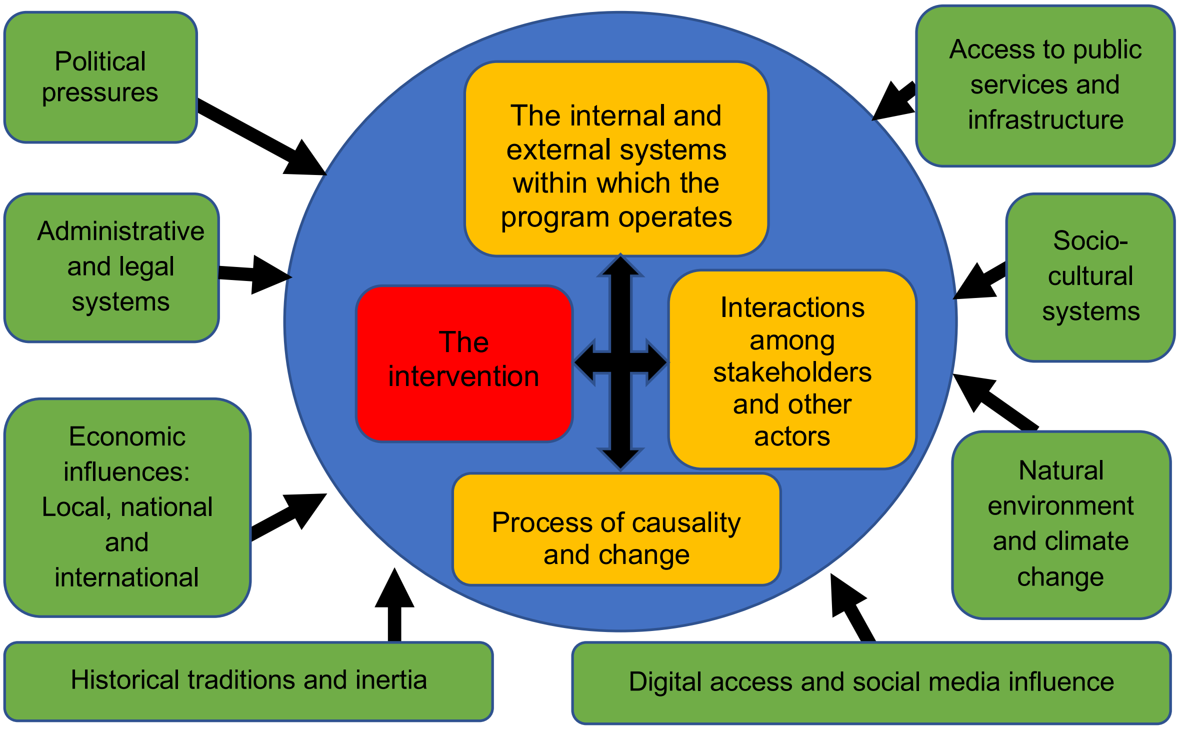
   While most managers and evaluators agree that projects operate in complex environments, in practice, very few evaluations address complexity. The failure to address complexity has serious implications for validity of the evaluation findings and recommendations. The complexity map identifies four dimensions of complexity that affect all evaluations:
   1. Complexity in the project
   2. The complex processes of interaction among the many different agencies involved in the finance, design, management and implementation and evaluation of the project
   3. The economic, political, socio-cultural, legal and administrative, demographic, climatic and other factors in the project environment that affect the project at all stages
   4. The processes of causality and change through which project outcomes and impacts are produced

   Part 1 of the 3ie blog on complexity-focused evaluation (see footnote 35 for the link) includes a complexity checklist that identifies a set of indicators that can be used to rate the level of complexity of each of the four dimensions identified in *Figure 8*. This provides managers, evaluators and other stakeholders with a clearer understanding of what is meant by complexity and which dimensions of a project are the most (and the least) complex. It also includes a basic five-step methodology for evaluating complexity, which is discussed in Part 1 of the complexity blog. The central message of the blog and these guidelines is that

[33] The complexity framework is described in the second part of the 3ie blog—Part 1 *Understanding real-world complexities for greater uptake of evaluation findings:* https://www.3ieimpact.org/blogs/understanding-real-world-complexities-greater-uptake-evaluation-findings, Part 2 *Building complexities into development evaluation:* https://3ieimpact.org/blogs/building-complexity-development-evaluations.

complexity must be taken into consideration at all stages of the evaluation; otherwise, the results of a process or impact evaluation can be misleading, often leading to an over-estimation of the project impact.

Figure 8: The complexity map


Political pressures
Administrative and legal systems
Economic influences: Local, national and international
Historical traditions and inertia
Digital access and social media influence
Natural environment and climate change
Socio-cultural systems
Access to public services and infrastructure
The internal and external systems within which the program operates
The intervention
Interactions among stakeholders and other actors
Process of causality and change

3. ***Experimental and quasi-experimental designs***. There are a wide-range of designs (see Gertler et al. 2011, Khandker et al. 2009, Bamberger and Mabry 2020, Chapter 12), with the choice depending on the characteristics of the project and time, budget and data constraints. In practice quasi-experimental designs (with matched comparison groups) are more widely used due to the relatively small number of situations where randomized assignment is possible. Designs are the strongest when data are collected both at the start and end of the project, but they can also be used in retrospective evaluations. These designs can include variants, such as multi-arm experimental designs, difference in difference, and interrupted time series among others (Massett, Shreshta and Juden 2021).34 While these designs are powerful, they are only appropriate for addressing certain kinds of evaluation questions (Bickman and Reich 2009) and for most process evaluations, they need to be complemented by other methods (see Design 4).

[34] This paper is a very useful resource for the present guidelines. Although it is part of a program on *Evaluating complex interventions*, many of the analytical approaches are directly applicable to the evaluation of different kinds of process and impact evaluation designs, including multicomponent interventions, portfolio interventions, interventions with long causal chains and system-level interventions. For the project summary on *Evaluating complex interventions*, visit: https://cedilprogramme.org/funded-projects/programme-of-work-1/evaluating-portfolio-interventions/ Cedil organized a webinar to discuss this paper, accessible at: https://cedilprogramme.org/blog/unpacking-complexity-in=international-development/

4. ***"RCT+."*** These are randomized control trials conducted within a mixed-methods framework. Qualitative methods are used to address issues that are difficult to capture with only quantitative survey data, or where it is helpful to use case studies to illustrate some of the important quantitative findings. These designs are also useful to track and explain outliers and findings that are inconsistent with the project theory (Bamberger, Tarsilla and Hesse-Biber 2016).[35]

5. ***Implementation research***. Implementation research has been defined as follows:

   Implementation research is the scientific inquiry into questions concerning implementation—the act of carrying out an intention into effect, which in health research can be policies, programs, or individual practices (collectively called interventions). –– Rutenberg and Heard 2018

   Originally developed in the health field, organizations such as 3ie are now applying the approach in other fields, such as education. There is considerable overlap between implementation research and other kinds of process evaluation, but in many cases, implementation research focuses more narrowly on the project and how it is implemented (e.g. using fidelity analysis), while some of the other process evaluation approaches have a broader focus on the wider context the project is implemented in. Some of the other approaches discussed in this document also try to study the informal socio-cultural and political processes that explain how implementation on the ground compares to the project design blueprint. Often, the differences can be important, particularly when they involve the exclusion of certain groups or significant reductions in efficiency. However, there is no single definition of implementation research.

6. ***Qualitative methods***. These designs use methods such as participant observation, in-depth interviews, group consultation methods, case studies to observe implementation processes, beneficiary behavior and attitudes, and observations of interactions among key actors. See Annex 2B for more detail.

7. ***Mechanism experiments***. Mechanism experiments are used to identify and test the underlying logic of the project design. RCTs are used to assess and compare the effects of different variations of policy and project interventions. One example is to test the "broken windows" theory that if an urban area has broken windows, graffiti and uncollected trash, this communicates the message that no one cares about the area and crime increases. A mechanism experiment might involve randomly selecting areas to leave cars with broken windows or broken bottles and trash, and compare crime rates with areas that are kept clean (Ludwig, Kling and Mullainathan 2011).

8. ***Contribution analysis*** uses a theory of change approach to define the "project story" as to how the project is intended to achieve its outcomes (Mayne 2012). All of the evidence is critically assessed to judge the credibility of the story—how outcomes are linked to the project. This is particularly useful for multi-donor

[35] Bamberger, M, Tarsilla, M, and Hesse-Biber, 2016. Why so many "rigorous" evaluation designs fail to identify unintended consequences of development programs: How mixed methods can help. Evaluation and program planning, 55: 155-162

projects to try to assess the contribution of a particular agency. Contribution analysis often only focuses on outcomes, but can also be used to critically assess the implementation design.

9. ***Case studies***. Case studies are used to provide more in-depth understanding of individuals or groups of interest to the study (Yin 2004 and 2012). They are also used to illustrate findings for different sectors of the target population. Case studies can be conducted in a few hours or take weeks or even years. When resources permit, a case study can cover a much longer period than what a survey, so they can examine how a process evolves or a family changes over time. In recent years, qualitative comparative analysis case studies have been used to assess outcomes when multiple characteristics of an individual or group interact to produce outcomes (configurational analysis) (Byrne and Ragin 2009). There are many potential applications for case studies process evaluation to provide in-depth descriptions of implementation processes, to illustrate how different groups respond to the project, or using Qualitative Comparative Analysis (Byrne 2009) to identify the configuration of factors in a case (household, organization etc) that are necessary for a project outcome to be achieved.

10. ***Realist evaluation***. This qualitative approach both assesses the influence of external factors, with emphasis on social control mechanisms, and tracks how project implementation changes in response to the interaction between project management and the different sectors of the target population (Pawson 2013). The realist evaluation framework is

    Context + Mechanism = Outcome

    “Because programs often work differently in different contexts and through different change mechanisms, they cannot simply be replicated from one context to another and automatically achieve the same results. Knowledge about what works, for whom, in what contexts, and how is, however, portable. Therefore one of the tasks of evaluation is to learn about contexts in which particular programs do and do not work, and what mechanisms are triggered by what programs in what contexts.” (Leeuw 2016). *Figure 9* illustrates how realist evaluation has been applied in many studies conducted to assess the effects of “naming and shaming” programs used to control the behavior of sex offenders who are released on parole. The realist framework examines the interactions among situational mechanisms that work at the macro level (agenda setting and diffusion of information), the local level (surveillance behavior by the local community) and the meso level (opportunity reduction and offender shame), which combine to reduce reoffending (Ashbury and Leeuw 2010).

Figure 9: Applying realist evaluation situational mechanisms to a "Naming and Shaming" project

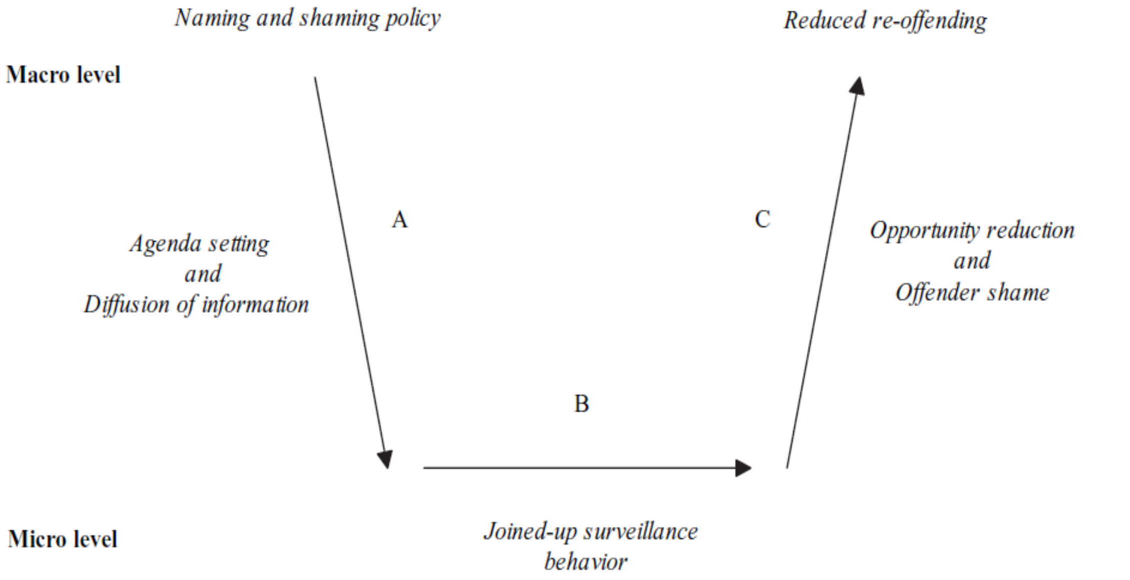


Figure 1. A basic model of mechanisms underlying 'naming and shaming" of sex offenders. A = Situational mechanism; B = Action-formation mechanism; C = Transformational mechanism. Source: Adapted from Coleman (1986); Hedström and Swedberg (1998).

11. ***Text analytics***. Many agencies have accumulated large volumes of PDF documents over the years. Policy statements, project design documents, monitoring and progress reports (among others) contain a great deal of valuable information on how projects were conceived, designed, implemented and evaluated. However, until recently, the format of these documents and huge number of pages made them virtually impossible to analyze. Now, text analytics makes it possible to scan millions of pages, and identify and track themes of interest. For example, all project documents could be scanned to identify the frequency of references to social exclusion, gender equality or climate change, or to review the strategies discussed with respect to implementation design. Documents from different agencies can also be combined in the analysis (Lourdusamy and Abraham 2018, Aggarwal 2018).

12. ***Big data and data analytics.*** Big data and data analytics are rapidly evolving and providing a widening range of data and data analytics tools and techniques that can be used for evaluations. *Table 10* lists some of the most widely used kinds of big data with examples of their applications in development evaluation. Data analytics also provides the tools for analyzing large data sets, including artificial intelligence and predictive analytics that broaden the scope of analysis. However, most evaluators are not yet familiar with these new tools and techniques, so the tools are still only used in a small portion of evaluations.[36] There are many potential applications for process evaluation that remain largely unexplored.

[36] A recent big data mapping survey conducted by 3ie found that most applications of big data in the development field were for research and only a small portion for impact evaluations.

Table 10: Big data sources and examples of their application in program evaluation

1. Geospatial analysis: Satellites and drones	Tracking population movements and the growth of informal and refugee settlements. Constructing natural experiments by matching project and control groups.
2. Social media (Facebook, Twitter etc.)	Sentiment analysis, identification of fake news and hate speech, feedback on social problems and concerns, organizational behavior
3. Radio call-in programs	Feedback on community problems and identifying sources of anti-refugee sentiment
4. Administrative records and secondary data (e.g. surveys conducted) from multiple agencies	Creating integrated data platforms to combine data sources from different agencies
5. Internet of things (mobility and biometric and mobility data from smart phones, and remote sensors)	The quantified self and quantified community, and monitoring use of community services (water and toilets)
6. Telecom call data records	Evaluating integration of refugees into host country

*13. **Systems analysis.*** While most development agencies recognize that the interactions among the large number of actors (stakeholders) involved in a development project can have a significant influence on the effectiveness with which a project is implemented and achieves the intended outcomes, most evaluations do not systematically assess these interactions and their effects. This is partly due to the methodological challenges in the analysis of these interactions. With the rapid evolution of big data and data analytics, it is becoming possible to apply many of the systems analysis tools required for these kinds of analysis. Some of the most promising tools include: systems mapping, systems dynamics, social network analysis, agent-based modelling and critical systems heuristics (boundary analysis) (Williams and Hummelbrunner 2011, Williams and Imam 2007, and Vaessen, Raimondo and Bamberger 2016). Many of these techniques have great potential for process evaluation, but to date, there is very little experience with their use.

Table 11 summarizes each of the 13 possible integrated process/impact evaluation designs and the potential level of applicability to each of the three process design scenarios. In some cases, the current use of a particular design may be quite low, but there may be a high potential for future use. All the designs must be adapted to the specific characteristics of each project, the context in which it operates, the kinds of data that are available and to time, budget and political constraints.[37]

[37] Many clients may have views on how they would like the evaluation to be organized, the questions they would like to see addressed and in some cases, issues they do not wish to be discussed or groups they do not wish to be included in the evaluation. For example, there are situations where the client does not wish the evaluation to include a control group or interview certain groups (such as NGOs who may be critical of the government). Mabry (Bamberger and Mabry 2020, chapter 8) discusses the many ways in which political influences can affect how the evaluation is designed, implemented and the results disseminated.

Table 11: Evaluation designs for each process evaluation scenario

	Process Evaluation Scenarios		
Possible evaluation designs	Retrospective/Ex-post evaluation	Pretest-posttest comparison group evaluation	Formative/ real-time evaluation
1. A mixed methods evaluation framework combines quantitative generalizability and statistical rigor with qualitative depth of analysis of behavioral change, and tracking of processes of change	*** It is recommended that researchers consider basing all impact/process evaluations on a mixed-methods framework. All quantitative impact evaluations require: (1) the use of qualitative data to permit a deeper analysis of, for example, informal implementation processes and (2) triangulation to strengthen the validity of key indicators		
2. Mapping complexity and assessing the influence of contextual factors and organizational arrangements on implementation and outcomes	** It is recommended that all evaluations include a complexity map. In some cases, this is also the framework for a more rigorous complexity-responsive evaluation	*** Longitudinal data provide more scope for rigorous complexity analysis	** This is often only used descriptively, but there is the possibility for more in-depth assessment of organizational arrangements and informal implementation processes
3. Experimental and quasi-experimental designs comparing groups with different project conditions (including fidelity to project design)	*** Applied retrospectively, often using recall and techniques such as propensity score matching	*** The longitudinal design permits the use of randomized control trials to rigorously test the influence of individual project components	* Not widely used, but could be useful with certain designs where the evaluation covers a longer time period
4. Experimental designs + (also known as RCT+): Integrates appropriate qualitative methods to create a mixed-methods experimental design	** Currently only used in a few experimental designs, but could potentially become a standard component	*** Currently only used in a few experimental designs, but could potentially become a standard component	* Most of these designs do not include an experimental component
5. Implementation research: Can include developing rating scales to assess project implementation fidelity	** Already used on a limited scale, but there is potential for expansion	*** Longitudinal data permit greater scope for application	* More limited application, but implementation rating scales can be used in most evaluations

	Process Evaluation Scenarios		
Possible evaluation designs	Retrospective/Ex-post evaluation	Pretest-posttest comparison group evaluation	Formative/ real-time evaluation
6. Qualitative methods to track implementation processes, beneficiary behavior and attitudes, and the influence of organizational and external factors	*** Incorporated into experimental (RCT+) designs	*** Incorporated into experimental + designs	*** Central element of formative evaluations
7. Mechanism experiments to identify and test the underlying logic of the project design; Using experimental designs to assess the effects of policy or program interventions	** Only possible to use natural experiments, which are weaker than RCTs	*** A potentially powerful tool that can, resources permitting, test a number of policy or project intervention options	* Less applicable
8. Contribution analysis: Uses a theory of change approach to define the "project story" as to how the project is intended to achieve its outcomes; Often, only focuses on outcomes, but can also be used to critically assess the implementation design	*** Contribution analysis can be used to strengthen all impact and process evaluation designs, and should be considered as a component of all mixed-method designs		
9. Case studies illustrating the experience of different individuals or sectors of the project population; Qualitative comparative analysis is a powerful tool for analysis of complexity	** Quite widely used to illustrate quantitative findings	** Quite widely used to illustrate and explain quantitative findings	*** Used extensively
10. Realist evaluation: Qualitative approach that both assesses the influence of external factors, with emphasis on social control mechanisms, and tracks how project implementation changes in response to the interaction between project management and different sectors of the target population	** Useful for addressing questions of who benefited from the project: when, how and why		*** A powerful approach for the analysis of how socio-cultural and political factors affect project implementation and outcomes

	Process Evaluation Scenarios		
Possible evaluation designs	Retrospective/Ex-post evaluation	Pretest-posttest comparison group evaluation	Formative/ real-time evaluation
11. Text analytics of project documents and other secondary data	** Starting to be used more widely and high potential applicability	** Starting to be used more widely and high potential applicability	* Not widely used, but potentially applicable
12. Designs drawing on the evolving big data and data analytics tools; Often involves the creation of an integrated data platform that can merge different kinds of data into a common metric, so that new kinds of comparative data analysis become possible	*** High potential applicability, but currently not widely used	*** High potential applicability, but currently not widely used	** Significant potential applicability, but currently not widely used
13. Systems analysis: Combines complexity-responsive evaluation with big data analytics	** Strong potential application, but designs are still being developed	*** Stronger applications due to the availability of longitudinal data	* More limited use depending on the kinds of data collected
Code: Level of applicability of the evaluation designs for each scenario: *** High potential applicability; ** moderate applicability; * low applicability or not often used			

3.5 The application of the different tools and techniques for integrated process/impact evaluations

Evaluation research methods that could be applied to process evaluation can be classified into four main groups. Annex 2 summarizes the most widely used research methods in each category with examples of how each method could be used in integrated process/impact evaluations.

3.5.1 Quantitative (QUANT) evaluation methods

QUANT methods normally work with largely numerical data—surveys, anthropometric and biometric data (height, weight, nutritional status), educational data on test scores, attendance and graduation rates, as well as income and consumption data, travel patterns, and so on. They can be used for descriptive, monitoring, process evaluation or experimental/quasi-experimental designs. In evaluation, a range of experimental and quasi-experimental designs are used to assess project performance or outcomes by constructing a counterfactual to compare changes in a project (experimental) group with a matched comparison group.

Potential benefits for integrated process/impact evaluations:

- QUANT methods use statistical analysis to compare variations in project outcomes (impacts) with variations in how the project was implemented, so as to assess the effects of implementation on outcomes. Two kinds of variations in project implementation can be assessed: (1) the fidelity of adhesion to the implementation protocol and (2) variations (intentional or unintentional) in the implementation strategy. Mechanism experiments (randomized comparisons of variations in policy and project implementation) are a potentially powerful, but currently under-used strategy for assessing effects of implementation modifications on outcomes and impacts.[38]
- By incorporating a comparison group, QUANT designs can control for the effects of external factors, such as improved economic conditions, the effects of other government policies and programs or the effects of other projects in the same areas supported by donors or NGOs.
- QUANT designs can also control for selection bias (e.g. when participants are not chosen randomly but are either self-selected or selected by project management, often to include those most likely to succeed).
- Quant designs yield more precise numerical information on the project population and the environment in which they operate. This can also be used for trend analysis.
- It is also possible to obtain more precise estimates of the size of outputs or outcomes, as well as more precise descriptive estimates of population size, production, general health status, and so on

[38] Examples of variations in implementation strategies that can be tested include: comparing the effectiveness of one and two week training programs, comparing the effectiveness of joint meetings of men and women participants to separate meetings for each sex, or varying the requirements to receive conditional cash transfers.

Potential limitations of an exclusive reliance on QUANT methods for integrated process/ impact evaluations:
Experimental designs require precise, quantifiable indicators of outputs and outcomes. The required data are often not available or cannot be quantified. Consequently, the analysis is forced to use proxy variables that do not precisely represent the key variables to be measured. This issue is called **construct validity**. For example, it is almost never possible to obtain accurate information on domestic violence, because the violence almost always occurs within the home and cannot be observed by the researchers. So studies of, for example, the effects of locating police units within the community on the rates of domestic violence must often rely on the number of reported complaints of domestic violence, responses of households to survey questions, or information from local health centers. All these sources significantly underestimate the incidence of domestic violence. Similarly, it is difficult to estimate household income for families in the informal sector, so many studies of household income rely on estimates of household consumption, which are easier to approximate, but often provide biased estimates of household income. Therefore, there are many cases where rigorous analytical methods are applied, but the indicators used in the analysis have major limitations.

- Most quantitative evaluations are designed to assess the impacts of a project on a limited set of project objectives/outcomes. They do not seek to identify the wide range of outcomes, planned and unplanned, that most interventions contribute to. This means that many evaluations potentially underestimate the effects (positive and negative) projects produce. Significant unintended negative outcomes are often ignored. Examples include (1) increased domestic violence resulting from projects that promote women's economic empowerment, (2) higher price of animal feed when previously freely available fodder is used to generate sustainable rural energy, and (3) the poorest businesses that cannot afford electricity to stay open after dark are put at a competitive disadvantage with the wealthier competitors, even in projects whose goal is to benefit the poorest and most vulnerable groups.
- As already mentioned, many QUANT evaluations are designed to assess project impacts when the effects of other factors are controlled for. Thus, many socio-cultural, economic and political factors are often excluded from the analysis. This seriously limits the ability of the evaluations to provide guidance on potential replication in other contexts.
- One or more of the four dimensions of complexity (see *Figure 8*) are often ignored. These excluded factors may significantly affect the evaluation of project outcomes.
- QUANT methods are very effective at comparing states (such as baseline and end-of-project), but they are less easy to apply for measuring dynamic processes, such as the project implementation process, or for studying behavioral change. This has serious implications for process evaluation, because when only QUANT methods are used, it is not possible to identify the informal social, cultural and political mechanisms that affect how project implementation actually takes place and how the implementation process changes over time. These variations, often very significant, are rarely found in any project reports or focus groups with project staff. This is an example of why this document strongly recommends the implementation of QUANT evaluations within a mixed-methods framework.

- Probably the greatest limitation of experimental methods is that in practice, they can only be applied in a very small fraction of evaluations. This is partly due to the cost and need for a high level of technical expertise, but also because there are many situations where a randomized design cannot be used for ethical, political or practical considerations.

3.5.2 Qualitative (QUAL) evaluation methods

QUAL methods are normally used with relatively small samples of individuals, communities, organizations or other units (e.g. case studies of policymaking processes). Each individual or group is studied intensively and sometimes, contact is maintained over long periods of time (some studies continue for a number of years). While individuals or groups are sometimes studied in isolation, usually the research tries to capture relationships of the particular subjects with the context they live or work in. While QUANT studies are usually concerned to ensure the representativity of the sample (often using random selection), there are a number of different ways to select samples for QUAL studies. Sometimes, a quota sample selection procedure is used. However, in ethnographic and other in-depth study methods, an *emergent* process is frequently used, whereby respondents are gradually identified over time as their relationship with the primary subjects becomes clear. In the following section on mixed-methods, we discuss the issues of representativity and the challenges of combining randomly selected samples with purposive or emergent QUAL samples.

Qualitative methods have several advantages for integrated process/ impact evaluations:

- QUAL studies can observe and describe the processes through which programs are implemented, and compare what actually happens in the real-world with intended project design and the expected processes and behavior. For example, a project goal may be to strengthen the participation of women in community decision-making, and project publications could report success as more women are attending community meeting. However, an observer attending the meeting might note that very few women speak or that when they do, male committee members do not take their comments.
- QUAL methods can also study how processes and behavior evolve over the lifetime of the project. Individuals and groups learn from their interactions with the project, and they often change their behavior and attitudes based on these experiences. Some groups may drop out, while others might encourage friends and neighbors to join. Clients can also pressure projects to introduce additional services or change how services are delivered (e.g. pressuring the project to hire more local language speakers, or convincing school feeding programs to also provide breakfast for younger siblings not originally covered by the program).
- Triangulation (comparing the consistency of data from different sources) is a powerful tool to improve the quality and validity of evidence used in the evaluation. This is discussed under mixed methods.
- In-depth interviews, participant observation and focus groups can often elicit more credible information from individuals who might not respond accurately to, or not understand, formal surveys.
- Individual, group and community behavior can be observed. For example, it is possible to observe informal community pressures that might discourage a teenage girl from taking vocational training classes outside the community, or encourage group members to get vaccinated or improve their diet.

- A major benefit is that it is possible to conduct longitudinal studies that trace how a program, or a community, evolve over long periods of time (e.g. starting two years before a project began and continuing for several years after the project closes). Longitudinal ethnographic studies have proved an effective way to track how microcredit programs could gradually change the economic and social status of women borrowers in the household and the community. Many of these changes can be slow, subtle, and difficult to capture through formal surveys.

Potential limitations of QUAL methods for integrated process/ impact evaluations:

- It is often difficult to generalize the findings of QUAL studies to the total research universe, because the sample of cases/subjects is frequently small and not selected to ensure representativity.
- The data are often more expensive and time-consuming to collect.
- More experienced interviewers and more expensive researchers are required for QUAL interviews than for enumerators of structured surveys.
- It can be more difficult to assess data quality, because the information is often collected in an unstructured way, depending in part on the interviewing style of each researcher and the particular circumstances surrounding interviews or observation.
- Unstructured interviews often generate large volumes of semi-structured or unstructured text, so the process of analysis is more time-consuming and complex. However, a wide range of qualitative data analysis and text analytics software is available to assist with the process.

3.5.3 Mixed methods and triangulation

Mixed methods provide the bridge that links QUANT and QUAL methods to combine the strengths of both approaches and compensate for the limitations of each (see earlier discussion of strengths and limitations of QUANT and QUAL). For most kinds of process evaluation, neither QUANT nor QUAL methods are fully adequate when used on their own (due to the limitations discussed earlier), so it is strongly recommended that all process evaluations use a mixed-methods framework.

There is an important distinction between a QUANT study that incorporates some QUAL data and a mixed-methods evaluation design framework. While most QUANT evaluations make some use of QUAL methods (e.g. focus groups, key informants, project visits), this is often done in a somewhat ad hoc way to explore issues that were identified during the analysis of the QUANT data; in contrast, a mixed-methods design involves a systematic strategy to integrate QUANT and QUAL tools and techniques at all stages of the evaluation.

Triangulation

Evaluation findings should not be based on unsubstantiated opinion or a few site visits where the evaluator may observe non-representative interactions, either innocent or planned. Triangulation is a technique used to compare estimates obtained from two or more estimates of the same variable using different methods of data collection. It is used for three main purposes:

1. To enhance validity
2. To create a more in-depth picture of a research problem
3. To interrogate different ways of understanding a research problem

OECD/DAC (2002:37) defines triangulation as: “The use of three or more theories, sources or types of information, or types of analysis to verify and substantiate an assessment. By combining multiple data sources, methods, analyses or theories, evaluators seek to overcome the bias that comes from single informants, single methods, single observers or single theory studies.” While the recommendation to compare three independent estimates of key variables derives from the nautical origin of the term (using three points to estimate distances), many mixed-method studies mainly rely on comparisons between only two independent sources, although a third source is preferable if available. Triangulation is crucial to strengthen validity and, as Bamberger and Mabry (2020: Chapter 14)[39] point out, it involves deliberate attempts to confirm, elaborate and disconfirm facts and interpretations through reference to the following:

- Multiple data sources
- Multiple methods of data collection
- Multiple evaluators and data collectors
- Repeated observations over time
- Multiple analytic perspectives
- Integrating the opinions and perspectives of all major stakeholders, including intended project beneficiaries and groups directly affected by the project, even if some of these groups were not included in the project design

Most often, triangulation helps validate research findings by checking that different methods or different observers of the same phenomenon produce the same results. It can also be used to interrogate inconsistencies between different data sources. The methodological framework used determines how the degree of overlap between methods is conceptualized. Researchers look for three types of triangulation: convergence, complementarity, and divergence. Convergence indicates there is a strong degree of overlap and accuracy between the data sets collected using different methods. Complementarity builds a richer picture of the research results by allowing the results from different methods to inform each other. Divergence presents a different set of challenges within the methods, and how it is interpreted depends on the conceptual framework for the research. Divergence can either indicate the methods or the results are flawed, or it may be treated as new data and analyzed to look for new insights.

Findings from a single data collection method can often be strengthened if they can be independently confirmed from two or more independent sources. This can be done in any of the following ways:

- Getting independent findings of change in variables (such as income, school enrolment, absentee rates, proportion of households using the village health center, etc.) from a variety of sources such as surveys, observations, focus groups, secondary data, etc.
- Comparing the findings through triangulation: If the findings that have appeared from using different methods are consistent, there can be greater confidence in the findings (*Figure 10*)
- If the findings are inconsistent, follow-up is required to determine the reasons and make adjustments to findings and conclusions (*Figure 11*)

[39] The American Institute for Research (AIR 2014) also provides a good framework.

Figure 10: Triangulation with converging findings of changes in household income[40]

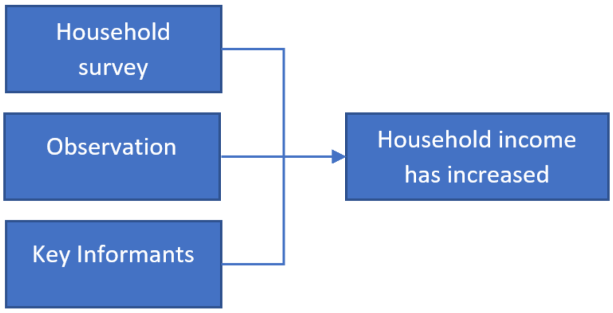
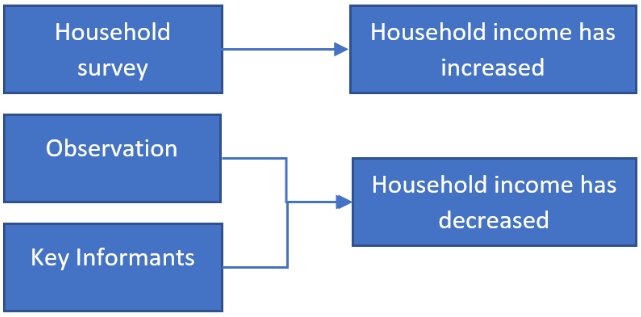


Figure 11: Triangulation with diverging estimates of changes in household income


Household survey
Household income has increased
Observation
Household income has decreased
Key Informants

The benefits of mixed methods for process evaluation:

- Combines the strengths of both QUANT and QUAL methods
- Combines depth of analysis provided by QUAL methods with the breadth and rigor provided by QUANT methods
- Combines the analysis of process and behavioral change with the accurate quantification for the total population
- Selecting an in-depth sample of project locations for the preparation of in-depth case studies

Challenges using mixed methods for process evaluation

- Requires integration of staff with QUANT and QUAL backgrounds, which often requires contracting more staff or consultants
- Requires a higher level of technical expertise and coordination
- Agencies familiar with QUANT methods may initially perceive mixed methods to be less professional and rigorous

3.5.4 Big data and data analytics

Big data and data analytics are rapidly evolving and provide a widening range of data and data analytics tools and techniques that can be used for evaluations. Annex 2.D lists some of the most widely used big data sources with examples of their applications in development evaluation. Data analytics also provides the tools for analyzing large data sets, including artificial intelligence and predictive analytics that broaden the scope of analysis. However, most evaluators are not yet familiar with these new tools and

[40] Figures 9 and 10 are taken from Bamberger et al. (2006:207).

techniques. As a result, these tools are still only used in a small proportion of evaluations. Also the recent 3ie big data mapping survey found that most applications of big data in the development field were for research and only a small portion for impact evaluations. Consequently, as evaluators' knowledge and comfort with big data are gradually strengthened, it will have a major impact on how evaluations are designed.

The potential benefits of big data for process evaluation:

- Reduces the cost and time of data collection and analysis, enabling significant increase in sample size, which makes it possible to conduct disaggregated analysis of different groups and sub-populations
- Broadens the types of data that can be collected and the types of measurement that are possible
- Increases the scope and coverage of data collection so that a range of contextual factors (local economic, political, socio-cultural and other factors) can be incorporated into the analysis
- Data can be collected over a longer period of time, beginning before the start of the project and continuing for several years after the project end, which is important for understanding trends and assessing sustainability
- It is possible to monitor processes over time, which is difficult to do with conventional evaluation tools
- It is possible to receive real-time feedback instead of having to wait for several months
- Supplies access to larger data sets, combined with real-time feedback, makes it possible to use systems analysis tools such as systems mapping, systems dynamics (modelling feedback loops among different stages of the project), social network analysis (useful for tracking patterns of interaction among different agencies and organizations over time) and principal agent analysis
- Provides real-time feedback on project implementation and the opinions and experience of different stakeholder groups (e.g. through social media analysis)

Challenges and limitations of big data for process evaluation:

- Many kinds of big data may not be accessible to a number of organizations for reasons of cost or difficulties of obtaining permission to use (e.g. phone company records)
- Many agencies lack the technical expertise to work with some kinds of big data
- Required information may not be available or there could be issues of data quality
- Many kinds of big data cannot be generated directly by the evaluation office, but must be generated and processed by other parts of the agency (usually the operations department). It may be difficult to convince operations to make the investment and set up the systems to collect the information. In many agencies, there would also be organizational and political constraints—the evaluation office may not be permitted to recommend to the operations department the kinds of data required for the evaluation, because these kinds of recommendations are sometimes considered to affect the independence and objectivity of the evaluations.

3.6 Observe, Divide and Surprise!

Spending time in the project area with beneficiaries is crucial for obtaining an understanding of the context and environment. The following are some practical tips and useful rules of thumb to bring to the field.

Expect surprises! Surprises always appear—in all evaluations. At least one. There will always be something that is different from what you have seen in any other project or evaluation. Some of the surprises could have been detected and anticipated earlier with better and more participatory planning during design, and some become apparent only during implementation. As evaluators, we need to expect the unexpected and look for the unanticipated. This is similar to detective work and has to do with being a good observer, asking open-ended questions, using your intuition and keeping your eyes and ears open.

It is useful to apply qualitative methods with open-ended questions to let the informants tell their story and not be confined to structured or semi-structured questionnaires or interview guides, because they often only provide the answers to your pre-defined questions. It is difficult to know what questions to ask when you do not know exactly which answers you are looking for. Make sure your focus and questions are not so narrow that you will miss out on information that can shed light on the underlying assumptions. If we only present pre-decided questions in a survey—which are, by definition, about already known factors—we miss the opportunity to discover what we do not know.

While conducting participatory rural appraisals (PRAs) in a village in Ethiopia, one of the authors had pre-drafted the framework for the seasonal calendar to save time. She had done numerous PRAs in 20 to 30 villages in other African countries prior to this study and thought she knew more or less what to expect (i.e. forgetting to still keep an open mind for surprises). She skipped the principle to *always start from scratch* and let the villagers draft the techniques; instead, she presented them with a pre-drafted calendar framework. It became very challenging to conduct the seasonal calendar exercise, because the information villagers provided about when they would prepare the land, when they would sow what kind of seed, when they would harvest and when they would sell their produce to be able to pay school fees, and so on simply did not add up. Toward the end of the exercise, one woman spoke up to explain that the drafted framework was wrong; they do not have 12 months in a year in Ethiopia. They have 13.

Direct observation is key, because one may be able to observe deviations from monitoring data or project reports, or any other unanticipated element. Direct observation is useful both to verify monitoring data (as with the community school teachers' attendance records in FATA and with the operational/non-operational SHGs in Ghana) and discover some of the project's unknown elements that might affect project implementation and results.

While doing a gender and poverty analysis in a Muslim area of Mozambique, we frequently observed that people kept pigs in their homesteads, while none of them reported having pigs when we did PRA exercises and interviews to map out their resources and food sources. Pigs were not mentioned at all in any of our project documents either, so we started asking about the pigs during interviews. The standard answer we got from all interviewees was that they did not own the pigs themselves, they were only looking after them for a neighbor. It turned out that people did keep and eat pigs in that area of Mozambique, and that pigs were, in fact, an important part of their food consumption, but it had previously been left out of all project planning and reporting because keeping and eating pork is a taboo for Muslims, and no one wanted to publicly admit the local common practice.

White’s observation of the non-operational SHGs in Andhra Pradesh is similar to one of the author’s observations of water user groups in Ethiopia and parent–teacher associations in Pakistan; many of the registered organizations cease to exist after a while, but they are still registered as operational groups in project documents. You may often not get a correct overview of the situation on the ground unless you go to the field, talk to people and observe.

Speak to “the angry man in the village.” This is a well-known secret most qualitative field workers are familiar with. If someone is angry or dissatisfied, it may turn out to be very useful to speak to them to find out why they are angry, because that will more often than not provide you with new and useful insights. White (2002:15 and White and Vajja 2008) give a good example from an evaluation of a rural livelihoods project in Andhra Pradesh in India, which included loans through women’s SHGs. White spoke to “an angry man in the village” who was upset about his unmarried daughter of 22 not getting access to a loan, because his wife had already received one. The conversation revealed that the project had regarded the loan to be for the household as a unit, not for each female individual family member, which meant that households with more than one eligible woman would only be able to take up one loan. This drove the project’s participation rates down significantly and was a very important element of implementation infidelity that had a negative impact on both the uptake and results of the project. It would not have been detected had White not taken the time to speak to the angry man.

Make surprise visits. Make sure to make some surprise visits to project sites that have not been prepared to receive you. In general, when planning the field visits, it is important that the evaluators, not the project staff, choose which locations to visit. Evaluators may apply certain site selection criteria to ensure representativity and avoid biases from project staff. Experience shows that more often than not, project staff are eager to show off the more successful parts of the project. But even after the formal planning, it might be a very good thing to just go off track and visit sites that have not been prepared for visits. This, of course, is only recommended in the cases that it may be done in an ethical manner. Schools that have been notified of visits from an evaluation team often have prepared songs and speeches for the evaluators, have cleaned, tidied and prepared the school and the kids have been drilled in the English alphabet. One of the authors went off track when assessing an education project in Pakistan to make a couple of surprise visits and found schools with only a handful of children playing alone in the school yard with no teacher present at all, as well as groups of children sitting with a teacher under a tree where school buildings were lacking. These observations would not have been made had she only stuck to the planned agenda, where all the pre-planned visited schools were very well prepared for the evaluator’s visit.

Be aware of who you are bringing to the field. It matters who comes with you to the field. If local people who travel with you are well known by the villagers (i.e. if they are a known politician or someone with power, they are the director of the project, etc.), their presence may influence what people dare to say or would want to say. The same goes for translators. While conducting a study in Nicaragua, our translator turned out to be the nephew of a known warlord who had previously raided the area and killed people. Our interviewees did not feel comfortable in our translator’s presence; luckily, someone notified us so that we could address the issue. If you bring the wrong people with you to the field, you will not get accurate information from the people you talk to.

Know the language or use professional translators. Take the time to brief your translator(s) well prior to start-up. Especially if they are not professionally trained translators, it is important to make sure that they understand their role well, they are not intimidating to the interviewees, they do not start to argue with the interviewees, and so on; they should understand that their role is simply to translate as accurately as possible exactly what is being said. They may convey their personal opinions (e.g. that man was lying, etc.) to the evaluator afterwards, when you are no longer with beneficiaries or other village people. Remember that the main point is to get authentic information from the users/beneficiaries themselves. Be aware that some may consider themselves to be superior to the community and as a result, they may have a somewhat strained relationship that might discourage some respondents from speaking freely.

Divide people and gather them. People will not talk freely if they are in front of people of power or people they do not trust. It is, therefore, recommended to make sure you talk to teachers away from headmasters and women away from men, poor people away from rich people, and so on—to let them speak freely and avoid control influence. However, this issue can be quite complex, because some researchers believe that vulnerable populations such as poor farmers, ethnic minorities, or some groups of women, may have more confidence to speak if there are some more articulate and experienced people from the community present. This is clearly an important question with no standard solution and where it is important to have a deep understanding of the community dynamics.

With PRAs, villagers are often divided into groups of women, groups of men and groups of leaders. Sometimes, it may also be a good idea to verify the information you have obtained from people individually and in separate groups when they are together. When conducting PRAs in various countries in Africa, we would gather the groups together toward the end of the day and present their findings to each other; this would often provide us with verifications of what people had reported (e.g. both men and women agreed that women worked 13 hours per day, while men worked 3 hours), and it would provide insightful discussions where difference of opinions had been conveyed.

4. Step-by-step—how to do it

These guidelines may be used in a flexible manner to incorporate process evaluation elements into larger program evaluations. They could be used for integrating process evaluation elements within a prospectively designed impact evaluation, and for post-hoc process evaluations of completed projects. They may be applied both prospectively and retrospectively.

More often than not, this is done by an external evaluation team not involved in project implementation. However, interaction and collaboration with people involved in implementation with regard to tracking inputs, process and outcome indicators is necessary.

While program theory cannot provide the precise statistical estimates of causality that can be obtained from experimental or strong quasi-experimental designs, theory models such as those presented in these guidelines provide a useful way to support or challenge evidence of causality obtained through mixed-methods evaluation design. Applying elements of theory-based process evaluation can provide useful indicators of probable linkages between the elements in the results chain.

The following is a step-by-step guide to how one may choose to plan and implement a theory-based process evaluation as part of an impact evaluation. It also summarizes the main points presented in the guidelines.

- **Understand the purpose of the evaluation, and understand and clarify the budget and timeline.** Budgeting for an evaluation consists of estimating the costs of the evaluation process. This is a useful framework for developing an evaluation budget. The budget will normally build on the *evaluation timeline/workplan*, where the different tasks are outlined and it is specified who on the evaluation team will do what. An evaluation budget would include:
    - *Consultants (and/or staff) time*: Number of days per evaluation staff member (this will build on the evaluation workplan) times (x) their different daily fees. Translators and enumerators time and fees should be included here, as well as time for planning/preparations/document reviews/ travel time/data collection time/time for analysis/time for report writing and time for dissemination of results.
    - *Travel expenses:*
        - Flights/train/bus/taxi expenses to and from evaluation destination
        - Taxi/car rentals etc. for domestic travel
        - Accommodation and Daily Subsistence Allowance for food and so on (times (x) number of days/times (x) number of staff
    - *Communication:* expenses related to communication, such as use of internet/mobile roaming/dissemination events/etc. Printing cost if the report is to be printed.
    - *Miscellaneous:* To ensure flexibility of the evaluation, it might be useful to calculate up to 10 percent of the budget to miscellaneous expenses, so you will be able to adjust for surprises and new discoveries that would need to be incorporated into the evaluation design and implementation.
- **Draw up an evaluation framework matrix.** Examples of evaluation framework matrices are found in *Table 7* and *Table 13* in these guidelines.
- **Read the project documents**, especially any progress reports, monitoring reports and mid-term reviews. Look for the **problem analysis** or any description of the diagnosis and prescribed treatment, the **theory of change** (program theory/logframe), and **logical gaps** and **underlying assumptions** in the theory of change. The usual starting point for putting together a program theory is the project documents. If there is a logical framework, then it embodies the program theory. However, it is unusual for a project document to make explicit all the underlying assumptions, although some of these may appear as risks.
- **Read a broader range of relevant literature,** such as existing evaluation studies and relevant academic literature, including anthropological ethnographies, economic studies, statistics, environmental analysis, and similar, to inform the evaluation design and identify logical gaps in the results chain. For socioeconomic and cultural context that might reveal unexpected underlying assumptions, any anthropological ethnography or study from the local area might prove useful. Identify some of the alternative hypothesis, alternative explanations for causal links and explanations concerning the expected outcomes and impacts. White (2002:15) states that in the Bangladeshi case, the identification of the mother-in-law effect came from reading anthropological literature and it led

the researchers to unpack the roster section of the questionnaire to identify the women (mothers) living with their mothers-in-law. The quantitative analysis was, thus, informed by important qualitative insight.

- **Create/recreate the theory of change and results chain** including all relevant contextual factors and underlying assumptions. The model should also recognize the likelihood of multiple causality and heterogeneity. Define operationally measurable input, output, outcome and process indicators. **Define indicators** with sufficient precision to permit them to be measured and quantified. **Define the time period** over which outcomes and impacts are expected to occur, and the intensity of inputs required to achieve outcomes. The time period selected to measure project effects can have significant impact on the estimated magnitude of the effects.
- **Run the proposed program theory by project managers.** Even if they had not thought it through explicitly before, they will have views on any such document that is produced. This exercise is a good opportunity to engage project managers, allowing them to influence evaluation design in beneficial ways.
- Determine the **main purpose of the evaluation,** the evaluation questions, information sources **and which quantitative and qualitative methods to apply based on your information need.** Using an evaluation framework/data matrix might be useful for this. An example of an evaluation framework can be found at the end of this chapter.
- **Recruit a multi-disciplinary evaluation team** that covers all necessary skills, sector skills, impact evaluation skills and evaluation experts experienced in qualitative field work.
- If possible, taking time and budget into account**, preliminary field work, including participatory analysis**, is an important part of evaluation design that can pick up unintended outcomes and other surprises, which can then be incorporated into the evaluation framework.
- **Go to the field, talk to the end users. Observe, divide and surprise!** It is important to talk to a wide variety of stakeholders, it is not sufficient to speak to ministry staff and/or project staff only. As pointed out by White (2002:15), there really is no substitute for spending time in the field yourself, and it is difficult to know how data can be sensibly analyzed without field exposure. The range of techniques goes from “development tourism” (spending a day or so in the field) through sufficient time in the field to use PRAs, which would require two to three days per village, to embedding an anthropologist in the project area for a longer period of time. The latter is hardly ever done, but could beneficially be used by longer-term studies, especially by collaborating with anthropology students who need to do longer-term field work for their dissertations. Nevertheless, even spending just a few days exposed to project implementation in a range of settings, preferably not just chosen by the project staff, will help both evaluation design and implementation. It can also be useful to visit non-project areas.
- **Apply the mixed-method approaches derived from your information need**. In addition to RCTs or quasi-experimental methods, do informal chats and interviews, (participatory) observations, use relevant PRA techniques, and so on. Use PRAs and other qualitative methods to identify causality. PRA has developed a variety of tools for working with community groups to identify

causality. Some of the time-related methods include timelines, trend analysis, historical transects (used to explore and represent the temporal dimension of people's reality[41] and seasonal diagrams) and relational methods(which include cause-effect diagrams, impact diagrams, systems diagrams, network diagrams, and process maps). All these techniques are based on working with stakeholders in facilitated discussions and exercises to construct maps, timelines, or causal chains defining the natural, political and sociocultural factors relevant to the program. Keep an open mind and let the beneficiaries fill in the forms. Make sure to divide people into appropriate groups, and ensure your translators and other people you bring with you to the field are neutral.

- **Data analysis and triangulation**: Define and combine all available evidence for inferring causality. Triangulate. The evaluation team will often have collected a number of different types of data that provide evidence on how the project has performed, what types of effects it has produced, and which groups have benefited the most and the least. Often, none of these sources of information are completely convincing when taken in isolation, but when they are combined and their consistency checked through triangulation, the evidence base becomes stronger.
- **Refine and adjust the program theory.** A program theory is never written in stone—it should be ready to adapt to surprises in the data.[42] Process evaluation results can be used to test theory (or parts of theory) and create new theory. This process encourages the use of theory to guide the planning and implementation of a process evaluation effort.
- **Produce a reader-friendly report that highlights the main findings and lessons learned** that may have implications for policy development and future programs.

[41] For example, farmers in a PRA session might be asked to draw pictures from different periods in the past, the present and the future of the seed varieties they use; access to water, forest cover, food availability and agricultural productivity (see e.g. Kumar 2002:143-48).

[42] As pointed out by White (2009:9), this approach may sound akin to data mining, but is in fact, quite different. The data miner knows what they are looking for and digs the data until they find it. The data analysis, on the other hand, is looking through the data allowing patterns, expected or unexpected, to emerge. See also Mukherjee et al. (1998).

Table 12: Step-by-step—how to do it

Stage	Steps
1. Preparing the evaluation	a. Understanding the purpose of the evaluation b. Understand and clarify the budget and timeline c. Draw up an evaluation framework matrix (examples in *Table 7* and *Table 13*) d. Read the project documents e. Assess the problem analysis and the evidence base and adequacy of the diagnosis f. Analyze the theory of change, look for underlying assumptions and logical gaps g. Read a broader range of relevant literature h. Diagnostic study, including where to do possible field visits i. Conduct stakeholder analysis j. Select a multidisciplinary evaluation team k. Develop collaborative relations with local research partners l. Create an advisory committee
2. Evaluation design	a. Create/recreate a theory of change and a results chain b. Define the information needs and prepare a data matrix c. Select the appropriate evaluation scenario d. Select the evaluation design and the data collection methods e. Integrate the evaluation design into a mixed-methods framework f. Discuss the proposed design with key stakeholders and the advisory committee g. Conduct an evaluability analysis h. Finalize the design, and define timelines and deliverables i. Review the possibility of including some of the new big data sources
3. Implementation	a. Define the organizational arrangements for the process evaluation b. Define coordination arrangements with the monitoring and management information systems of the implementing agency and relevant partners c. Put in place quality control mechanisms, including triangulation d. Conduct periodic review and revision of the theory of change
4. Analysis, reporting and dissemination	a. Produce a user-friendly report highlighting the main findings and lessons b. Ensure the report is accessible to different groups c. Develop mechanisms to consult with and obtain feedback from all key stakeholders d. Consider webinars, conferences and other dissemination events
5. Promoting the use of the findings and lessons	a. Consider a management response matrix b. Consider a follow-up to check on utilization

Table 13: Example of an evaluation framework matrix

Evaluation questions	Information needed	Sources	Methods applied	Limitations	Potential answers
Identify the specific evaluation questions. Ensure that all key terms are defined. Each major evaluation question should be addressed in a separate row.	What information do you need to be able to respond to the evaluation question? Specify what information is needed to respond to the question	Where will you find the information needed? Specify the different information sources (e.g. specific stakeholders with relevant knowledge, specific reports or publications, national statistics, monitoring data, etc.)	How will the team answer each evaluation question? Describe strategies and methods for collecting the required information or data (e.g. random sampling, case studies, key informant interviews, focus groups, questionnaires, beneficiary surveys, benchmarking to best practices, use of existing data bases, document review, etc.) and the analytical techniques to be used (e.g. regression analysis, cost benefit analysis, sensitivity analysis, modelling, descriptive analysis, content analysis, case study summaries, etc.)	What are the design's limitations and how will they affect the evaluation? For example: Questionable data quality and/or reliability Inability to access certain types of data or obtain data covering a certain time frame Unavailability of some key informants Security classification or confidentiality restrictions Ethical dilemmas Be sure to address how these limitations will affect the evaluation.	What kind of answer(s) will this part of the evaluation likely provide? Describe what the evaluation team can likely say. Draw on preliminary results for illustrative purposes, if helpful. Ensure that the proposed answer addresses the evaluation question in column one.
Evaluation question 1					
Evaluation question 2					
Evaluation question 3					
Evaluation question 4					

Other publications in the 3ie working paper series

The following papers are available from http://3ieimpact.org/evidence-hub/publications/working-papers

Promoting women's groups for facilitating market linkages in Bihar, India, 3ie Working Paper 49. Kochar, A, Tripathi, S, Rathinam, F, Sengupta, P and Dubey, P, 2021.

What stimulates the demand for grid-based electrification in low-and middle-income countries? 3ie Working Paper 48. Lane, C, Prasad, SK and Glandon, D, 2021.

Improving delivery and impacts of pro-poor programmes, 3ie Working Paper 47. Barooah, B, Jain, C, Kejriwal, K, Sengupta, P, Shah, P, Shah, R and Jain, S, 2021.

Understanding India's self-help groups: an organisational anatomy of functionality in a district in Madhya Pradesh, 3ie Working Paper 46. Bhanjdeo, A, Narain, N, Sheth, S and Walton, M. 2021.

Women's economic status and son preference: empirical evidence from private school enrolment in India, 3ie Working Paper 45. Gupta, R, Jain, S, Kochar, A, Nagabhushana, C, Sarkar, R, Shah, R and Singh, G, 2021.

Understanding barriers to and facilitators of latrine use in rural India, 3ie Working Paper 44. Jones, R and Lane, C, 2021.

Quality improvement approaches to enhance Iron and Folic Acid Supplementation in antenatal care in Uganda, 3ie Working Paper 43. Tetui, M, et al, 2021.

Assessing bottlenecks within Iron and Folic Acid Supplementation Delivery in Uganda: a workshop report, 3ie Working Paper 42. Agabiirwe, C, Luwangula, A, Tumwesigye, N, Michaud-Letourneau, I, Rwegyema, T, Riese, S, McGough, L, Muhwezi, A. 2021.

Literature review on selected factors influencing Iron Folic Acid Supplementation in Kenya and East Africa, 3ie Working Paper 41. Njoroge, B, Mwangi, A, Okoth, A, Wakadha, C, Obwao, L, Amusala, B, Muithya, M, Waswa, V, Mwendwa, D, Salee, E, Njeri, T and Katuto, M, 2021.

The policies that empower women: empirical evidence from India's National Rural Livelihoods Project, 3ie Working Paper 40. Kochar, A, Nagabhushana, C, Sarkar, R, Shah, R and Singh, G, 2021.

Assessing bottlenecks within Iron and Folic Acid Supplementation Delivery in Kenya: a workshop report, 3ie Working Paper 39. Njoroge, BM, Mwangi, AM and Letourneau, IM, 2020.

Mapping implementation research on nutrition-specific interventions in India. 3ie Working Paper 38. Tripathi, S, Sengupta, P, Das, A, Gaarder, M and Bhattacharya, U, 2020.

The impact of development aid on organised violence: a systematic assessment, 3ie Working Paper 37. Zürcher, C, 2020.

The current and potential role of self-help group federations in India, 3ie Working paper 36. Barooah, B, Narayanan, R and Balakrishnan, S, 2020.

How effective are group-based livelihoods programmes in improving the lives of poor people? A synthesis of recent evidence. 3ie Working Paper 35. Barooah, B, Chinoy, SL, Bagai, A, Dubey, P, Sarkar, R, Bansal, T and Siddiqui, Z, 2020.

Social protection: a synthesis of evidence and lessons from 3ie evidence-supported impact evaluations, 3ie Working Paper 34. Tripathi, S, Kingra, KJ, Rathinam, F, Tyrrell, T and Gaarder, M, 2019.

Transparency and accountability in the extractives sector: a synthesis of what works and what does not, 3ie Working Paper 33. Rathinam, F, Cardoz, P, Siddiqui, Z and Gaarder, M, 2019.

Integrating impact evaluation and implementation research to accelerate evidence-informed action, 3ie Working Paper 32. Rutenberg, N and Heard, AC, 2018.

Synthesis of impact evaluations of the World Food Programme's nutrition interventions in humanitarian settings in the Sahel, 3ie Working Paper 31. Kaul, T, Husain, S, Tyrell, T and Gaarder, M, 2018.

Community-driven development: does it build social cohesion or infrastructure? A mixed-method evidence synthesis, 3ie Working Paper 30 White, H, Menon, R and Waddington, H, 2018.

Evaluating advocacy: an exploration of evidence and tools to understand what works and why. 3ie Working Paper 29. Naeve, K, Fischer-Mackey, J, Puri, J, Bhatia, R and Yegbemey, R, 2017.

3ie evidence gap maps: a starting point for strategic evidence production and use, 3ie Working Paper 28. Snilstveit, B, Bhatia, R, Rankin, K and Leach, B (2017)

Examining the evidence on the effectiveness of India's rural employment guarantee act, *3ie Working Paper 27*. Bhatia, R, Chinoy, SL, Kaushish, B, Puri, J, Chahar, VS and Waddington, H (2016)

Power calculation for causal inference in social science: sample size and minimum detectable effect determination, 3ie Working Paper 26. Djimeu, EW and Houndolo, DG (2016)

Evaluations with impact: decision-focused impact evaluation as a practical policymaking tool, 3ie Working Paper 25. Shah, NB, Wang, P, Fraker, A and Gastfriend, D (2015)

Impact evaluation and policy decisions: where are we? A Latin American think-tank perspective, 3ie Working Paper 24. Baanante, MJ and Valdivia, LA (2015)

What methods may be used in impact evaluations of humanitarian assistance? 3ie Working Paper 22. Puri, J, Aladysheva, A, Iversen, V, Ghorpade, Y and Brück, T (2014)

Impact evaluation of development programmes: experiences from Viet Nam, 3ie Working Paper 21. Nguyen Viet Cuong (2014)

Quality education for all children? What works in education in developing countries, 3ie Working Paper 20. Krishnaratne, S, White, H and Carpenter, E (2013)

Promoting commitment to evaluate, 3ie Working Paper 19. Székely, M (2013)

Building on what works: commitment to evaluation (c2e) indicator, 3ie Working Paper 18. Levine, CJ and Chapoy, C (2013)

From impact evaluations to paradigm shift: A case study of the Buenos Aires Ciudadanía Porteña conditional cash transfer programme, *3ie Working Paper 17*. Agosto, G, Nuñez, E, Citarroni, H, Briasco, I and Garcette, N (2013)

Validating one of the world's largest conditional cash transfer programmes: A case study on how an impact evaluation of Brazil's Bolsa Família Programme helped silence its critics and improve policy, *3ie Working Paper 16*. Langou, GD and Forteza, P (2012)

Addressing attribution of cause and effect in small n impact evaluations: towards an integrated framework, 3ie Working Paper 15. White, H and Phillips, D (2012)

Behind the scenes: managing and conducting large scale impact evaluations in Colombia, *3ie Working Paper 14*. Briceño, B, Cuesta, L and Attanasio, O (2011)

Can we obtain the required rigour without randomisation? 3ie Working Paper 13. Hughes, K and Hutchings, C (2011)

Sound expectations: from impact evaluations to policy change, 3ie Working Paper 12. Weyrauch, V and Langou, GD (2011)

A can of worms? Implications of rigorous impact evaluations for development agencies, 3ie Working Paper 11. Roetman, E (2011)

Conducting influential impact evaluations in China: the experience of the Rural Education Action Project, *3ie Working Paper 10*. Boswell, M, Rozelle, S, Zhang, L, Liu, C, Luo, R and Shi, Y (2011)

An introduction to the use of randomised control trials to evaluate development interventions, *3ie Working Paper 9*. White, H (2011)

Institutionalisation of government evaluation: balancing trade-offs, *3ie Working Paper 8.* Gaarder, M and Briceño, B (2010)

Impact evaluation and interventions to address climate change: a scoping study, *3ie Working Paper 7*. Snilstveit, B and Prowse, M (2010)

A checklist for the reporting of randomised control trials of social and economic policy interventions in developing countries, *3ie Working Paper 6.* Bose, R (2010)

Impact evaluation in the post-disaster setting, 3ie Working Paper 5. Buttenheim, A (2009)

Designing impact evaluations: different perspectives, contributions, 3ie Working Paper 4. Chambers, R, Karlan, D, Ravallion, M and Rogers, P (2009) [Also available in Spanish, French and Chinese]

Theory-based impact evaluation, *3ie Working Paper 3*. White, H (2009) [Also available in French and Chinese]

Better evidence for a better world, *3ie Working Paper 2.* Lipsey, MW (ed.) and Noonan, E (2009)

Some reflections on current debates in impact evaluation, *3ie Working Paper 1*. White, H (2009)

Working Paper Series

International Initiative for Impact Evaluation
Unit no. 306,3rd Floor, Rectangle-1
D-4, Saket District Centre
New Delhi – 110017
India

3ie@3ieimpact.org
Tel: +91 11 4989 4444

www.3ieimpact.org