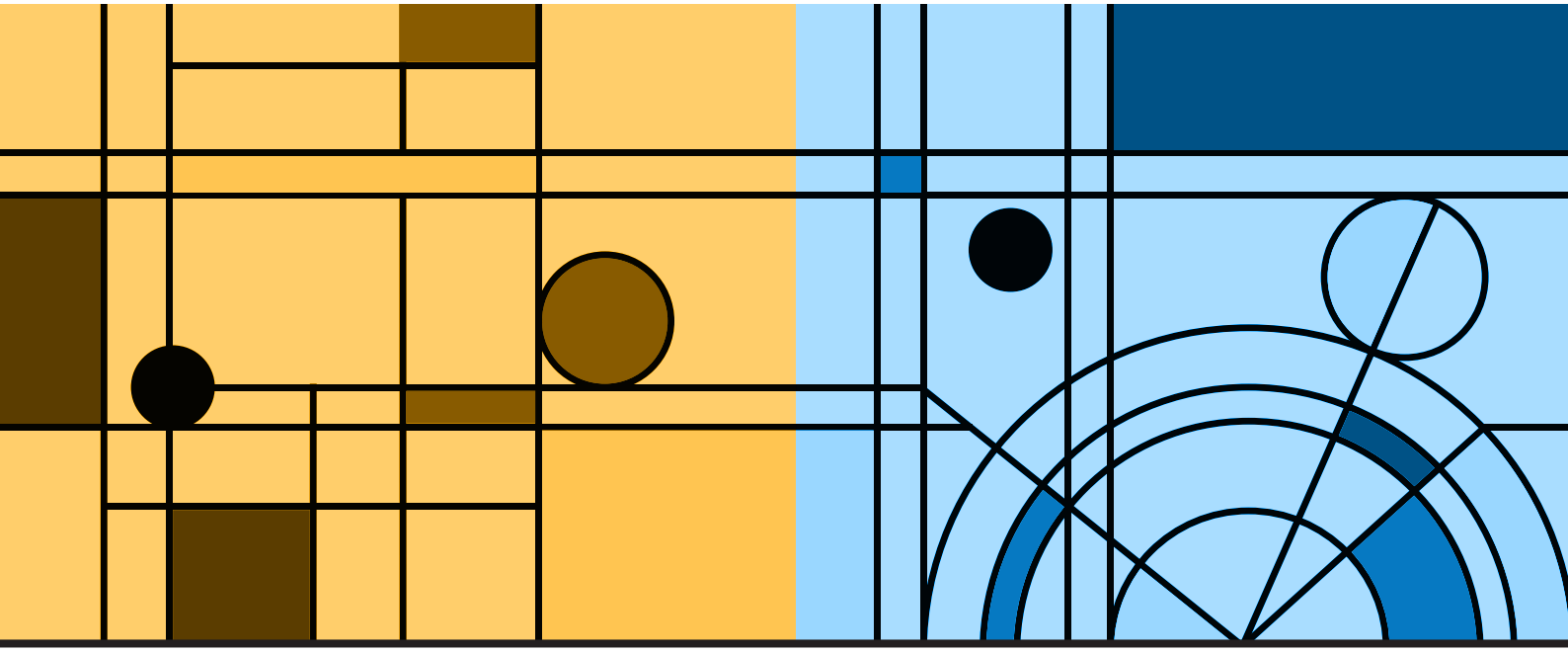
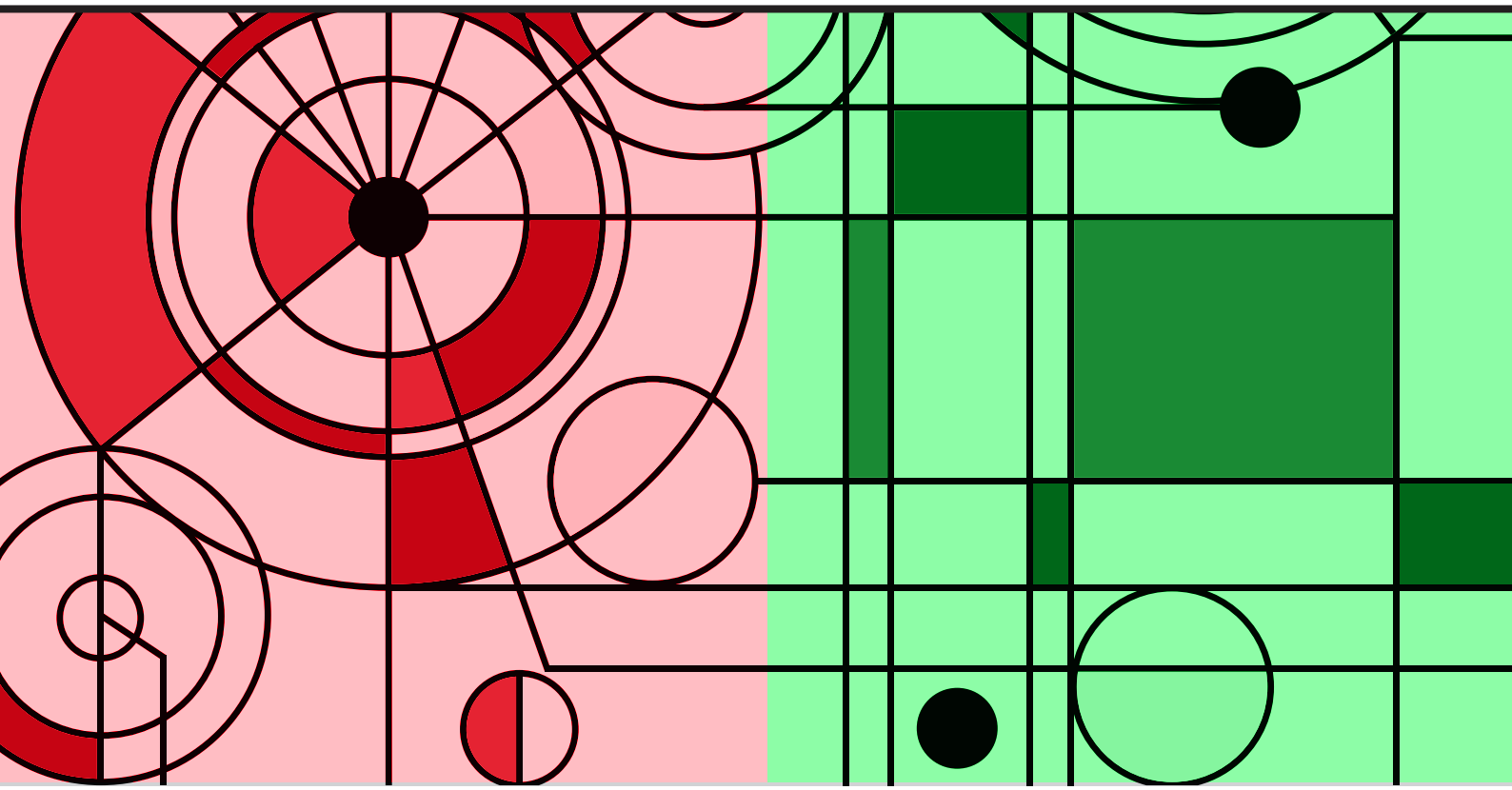
European Commission



EVALUATION HANDBOOK



International Partnerships

EVALUATION HANDBOOK

July 2024

Manuscript completed in July 2024
This document should not be considered as representative of the European Commission's official position
Luxembourg: Publications Office of the European Union, 2024

PDF ISBN 978-92-68-19841-4 ISSN 2529-3338 doi: 10.2841/340793 MN-02-24-847-EN-N

ACKNOWLEDGEMENTS

This Evaluation Handbook was prepared by a task force composed of staff from the Directorate-General for International Partnerships (DG INTPA), Unit D4, including its former Evaluation Support Service (ESS), under the guidance of a management board from DG INTPA and the Service for Foreign Policy Instruments (FPI).

The handbook was authored and edited by the task force members along with contributions from external writers, with valuable input from colleagues across DG INTPA, the Directorate-General for Neighbourhood and Enlargement Negotiations (DG NEAR), and FPI, including its service on Monitoring, Evaluation, Learning, and Design in External Action (meldea).

The lead authors for the various chapters of this handbook were (in alphabetical order) Hur Hassnain, Karen McHugh, and Marco Lorenzoni (Chapter 1); Valentin Alvarez (Chapter 2); Anna Maria Augustyn, Rick Davies, and Patricia Rogers (Chapter 3); and Margie Buchanan-Smith and Hur Hassnain (Chapter 4).

Sincere thanks to Howard White, Centre of Excellence for Development Impact and Learning (CEDIL), and Florencia Tateossian, UN Women, for their invaluable peer review, which significantly enriched this handbook. Special appreciation goes to Nita Congress for her exceptional editing and design contributions.

Contents

List of boxes

List of figures

List of tables

Foreword

The 2024 edition of the Evaluation Handbook represents a revised version of the Methodology Guidance published in 2006. It incorporates contemporary evaluation literature, international best practices and standards, and leverages the wealth of experience and good practices. Extensive input from consultations with various staff members of the Directorate-General for International Partnerships (DG INTPA), the Directorate-General for Neighbourhood and Enlargement Negotiations (DG NEAR) and the Service for Foreign Policy Instruments (FPI) has ensured alignment with DG INTPA's and FPI's approach, processes and with the European Commission's Better Regulation Guidelines. The handbook is available electronically, enabling periodic revisions and updates.

Abbreviations and acronyms

CRIS	Common RELEX Information System
DAC	Development Assistance Committee
EC	European Commission
EEAS	European External Action Service
ESS	Evaluation Support Service
EU	European Union
EWP	Evaluation Work Programme
GDPR	General Data Protection Regulation
GERF	Global Europe Results Framework
GIS	Geographic Information System
ICT	Information and Communication Technology
InCA	Intercultural Approach
DG INTPA	Directorate-General for International Partnerships
DG NEAR	Directorate-General for Neighbourhood and Enlargement Negotiations
FPI	Service for Foreign Policy Instruments
GEM	Global Elimination Methodology
IPA	Instrument for Pre-accession Assistance
IT	Information Technology
JRC	Joint Research Centre
M&E	Monitoring and Evaluation
MIP	Multi-annual Indicative Programming
MO	Modus Operandi
NGO	Non-Governmental Organisation
OECD	Organisation for Economic Co-operation and Development
QAG	Quality Assessment Grid
QCA	Qualitative Comparative Analysis
RELEX	External Relations
RIP	Regional Indicative Programming
ROM	Results-Oriented Monitoring
SMART	Specific, Measurable, Attainable, Relevant and Time-bound
ToR	Terms of Reference

Introduction

Evaluation matters. In an increasingly complex and challenging environment for international development and partnership, the importance of demonstrating results and learning from the European Commission's (EC's) external assistance is critical. Evaluation is a key learning tool for the EC to understand not only **what works** and **what does not**, but **why** and **under what circumstances**. Evaluations generate knowledge and produce evidence which we use to improve the way we engage with our partners and enhance the impact of our development cooperation.

As one of the world's largest donors[1], the European Union (EU) is a leading force in demonstrating the value of rigorous evaluation. The EU is committed to embedding an '**evaluate first**' culture (EC, 2013), premised on using evaluation as an indispensable tool to inform its choices and decisions with the best available evidence and thereby improve its strategies and practices.

This handbook is a revision of the Methodology Guidance published in 2006 and reflects the significant progress made in the field of evaluation over the past two decades. **The handbook is a living document** that will be adapted over time to reflect evolving practice and needs. It will also be made available as a web-based platform.

Target audience

This handbook is primarily aimed at:

- evaluation managers in EU delegations and at headquarters;
- monitoring and evaluation focal points in EU delegations;
- external evaluation teams and contractors;

[1] In 2022, official development assistance by member countries of the Development Assistance Committee (DAC) of the Organisation for Economic Co-operation and Development amounted to USD 204.0 billion, of which USD 91.6 billion was provided by the 20 DAC countries that are EU Members; this represented an increase of 18.6 per cent in real terms compared to 2021, and 0.57 per cent of their combined gross national income.

- experts and stakeholders who have been asked to serve on evaluation reference groups;
- partners and other stakeholders who wish to understand how evaluation is conducted in external action at the European Commission.

The handbook is intended as a reference for other evaluation stakeholders as well – including but not limited to other donors, partner governments, the private sector and civil society – and the professional evaluation community.

How to use this handbook

This handbook aims to describe how evaluation is performed at the Directorate-General for International Partnerships (DG INTPA) and the Service for Foreign Policy Instruments (FPI)[2]. To ensure maximum user-friendliness, the information presented in this handbook is of a 'hands-on' nature: **what has to be done and how to go about doing it**. Hyperlinks are provided to connect to (i) other parts of the handbook where more detailed explanations or related information can be found; (ii) online guidance, forms and templates to be used; and (iii) the OPSYS web portal (see below) and its evaluation section. The handbook can thus be read sequentially or as different stand-alone parts.

> **NOTE:** *Throughout, EC internal weblinks are in grey italic; all other links are in green.*

The handbook consists of four chapters and one annex (on budget support) and a comprehensive glossary:

- Chapter 1: Role of evaluation in DG INTPA and FPI introduces the main concepts; describes the various uses and types of evaluations and their primary users; and situates these within the intervention cycle.
- Chapter 2: Managing an evaluation provides hands-on practical guidance for managing evaluations through the six-phase evaluation process[3]. It is divided into eight sections, each correlating to an evaluation phase along with (i) an introductory section describing these phases and the main evaluation stakeholders and (ii) a section on the cross-cutting issue of quality assurance.
- Chapter 3: Approaches, methods and tools supplements the how-to guidance of Chapter 2 with detailed explanations, examples, rationales and techniques about topics mentioned elsewhere in the handbook. In particular, it provides a 'deep dive' into evaluation approaches, evaluation criteria and questions, evaluation methodologies, data collection tools and management, and data analysis.
- Chapter 4: Ethics in evaluation underlies all of the material presented in the preceding chapters. Conducting evaluations in an ethical manner is imperative. This chapter discusses fundamental ethical principles and considerations in evaluation and presents proactive, hands-on guidance for specific evaluation aspects and contexts.

Evaluation tools and resources

GUIDELINES AND HANDBOOKS

The handbook is built on the Better Regulation Guidelines (EC, 2021a) and Better Regulation Toolbox (EC, 2023) and other relevant EC guidance such as the ROM Guidelines, the *Evaluation Terms of Reference templates and guidance* and other methodological notes issued by DG INTPA and FPI. In addition, the handbook takes stock of information from external (non-EC) sources; these are all fully referenced and/or hyperlinked where possible in the document for further reading.

[2] Note that although various sections of this handbook will be relevant for evaluations conducted by other directorates and units of the EC, as well as for other organisations, its main focus is on how evaluations are conducted at DG INTPA and FPI. It contains basic guidance on evaluation in general and provides useful references to external sources where more information can be found.

[3] Although this handbook specifically targets intervention-level and strategic evaluations, it is equally relevant for other types of evaluations. See Section 1.2.

OPSYS

OPSYS is a web-based information technology (IT) ecosystem for EC staff and implementing partners. At full functionality, it will incorporate and replace all pre-existing IT systems used in managing the EU external cooperation portfolio.

THE EVALUATION WIKI

The Evaluation Wiki has been designed as a one-stop shop for all evaluation-related guidance material and documents. It is currently articulated in four main sections:

- Evaluation methodological guidance
- Evaluation help desk support
- Monitoring and evaluation focal points
- Other useful links

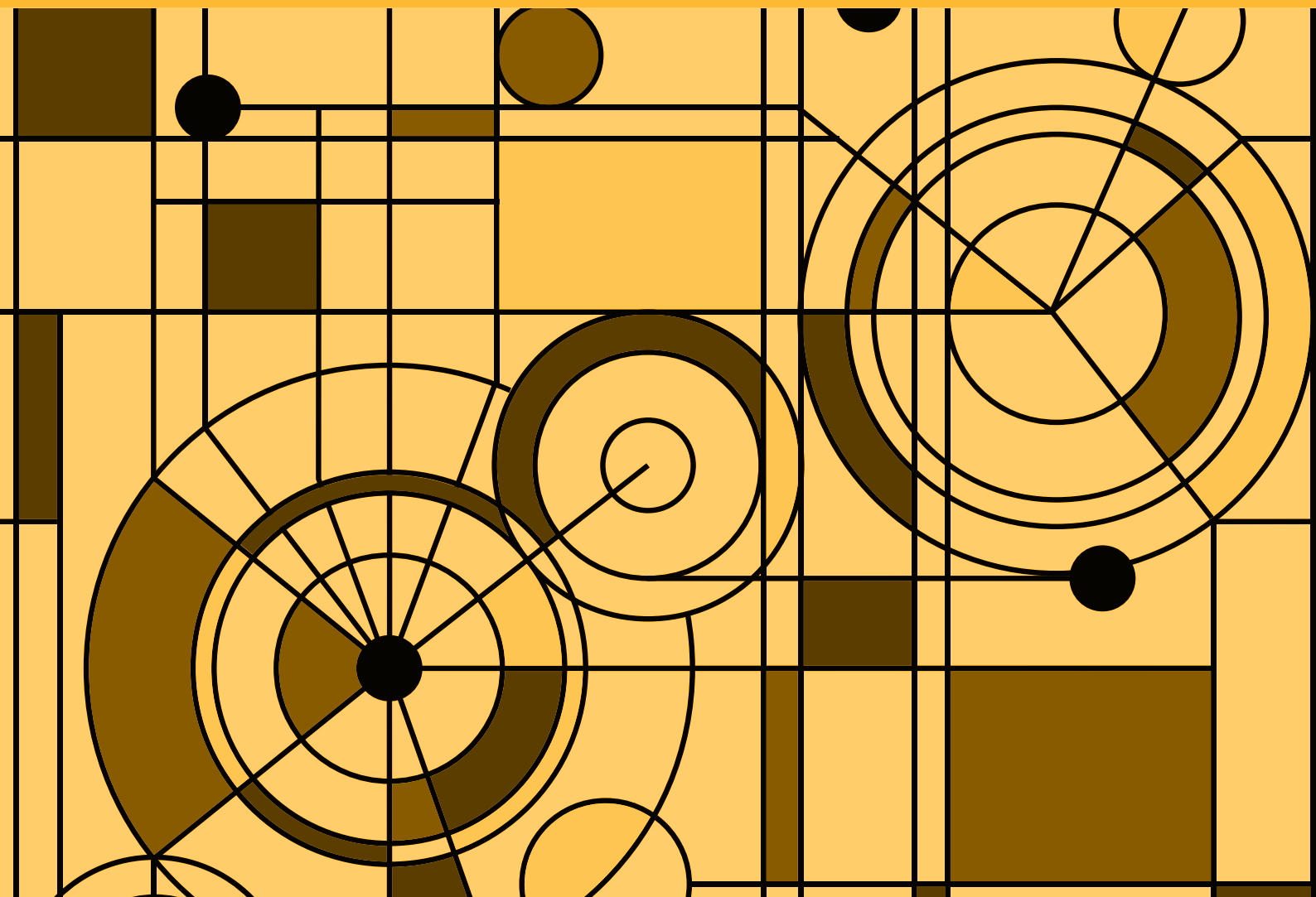
CAPACITY4DEV

Capacity4dev is the EC's knowledge-sharing platform for international cooperation and development. The platform hosts various collaborative online workspaces and communities for the exchange of knowledge on evaluation, including the following:

- Evaluation methodological approach public group
- Public Group on Design, Monitoring & Evaluation
- The Monitoring and Evaluation Focal Point Network, made up of staff members who manage, plan or advise on evaluations conducted in EU delegations and headquarters units (restricted group)
- INTPA/ESS Initiatives, a series of evaluation-related initiatives launched by the former DG INTPA Evaluation Support Service team

1

Role of evaluation in DG INTPA and FPI



What is this chapter about?

This chapter introduces the main concepts related in evaluation as practised in the Directorate-General for International Partnerships and the Service for Foreign Policy Instruments. It describes the various uses and types of evaluations and their primary users, and situates these within the intervention cycle.

How will this help you in your work?

This chapter explains the various elements of an evaluation – its purpose, its triggers, its types and timing, its key stakeholders – and tells you how these all fit together in planning an evaluation.

For definitions of key terms used in this handbook, refer to the glossary.

The European Union (EU) addresses shared global responsibilities through international partnerships that uphold and promote European values and interests and contribute to world peace and prosperity. As part of the EU's external relations, the Directorate-General for International Partnerships (DG INTPA) is at the forefront of these partnership-based efforts aimed at contributing to sustainable development, poverty eradication and the promotion of peace and the protection of human rights. The Service for Foreign Policy Instruments (FPI) puts EU foreign policy into action, fast and flexibly, in a policy-driven and integrated approach, and act as first responder to foreign policy needs and opportunities. It does so by helping countries cope with crises and maintain peace and security, by observing elections to support democracy and the rule of law, and by building alliances and leveraging the EU's influence in the world.

To build stronger, effective cooperation with partner countries, the EU relies on quality evidence generated through the systematic use of robust evaluation findings and recommendations. To this end, it has developed and drives a **strong culture of accountability and learning**, demonstrating results, and using evidence to enhance policies and practice. This commitment is set forth in the cornerstone EU Evaluation Policy document Evaluation Matters (EEAS and EC, 2014).

Supporting this commitment, the European Commission (EC) makes evaluation a central part of international partnerships, rooted in an awareness of its importance in the EU institutional culture. The EC's Better Regulation Guidelines (EC, 2021a) recognise evaluation as supporting decision-making and contributing to strategic planning and the design of future interventions. The EC applies the 'evaluate first' principle (EC, 2013) to ensure that any policy decisions take into account lessons from past EU action. Management and staff in EU delegations and at headquarters are continually encouraged to make extensive use of evaluation findings to better support the efforts of partner countries to eradicate poverty, improve governance and attain sustainable growth.

1.1 What is evaluation?

Evaluation is commonly understood as the 'systematic and objective assessment of a planned, ongoing or completed intervention, its design, implementation and results' (OECD DAC, 2023). Within the EC, evaluation is used to assess the performance of a strategy, policy, instrument, modality, intervention or group of interventions.

> **NOTE:** *Consistent with current practice, this handbook uses 'intervention' generically to mean 'project' and/or 'programme'.*

The EU's Evaluation Policy – set out in Evaluation Matters – governs the evaluation functions and practices of the EC's external actions. Principles governing the evaluation of EU international cooperation and development policies and activities are set out in the EC's Better Regulation Guidelines (EC, 2021a), which in turn explicate the reforms set out in Strengthening the Foundations of Smart Regulation – Improving Evaluation (EC, 2013). The methodology deriving from the EU Evaluation Policy prioritises results in evaluating EC external action.

1.1.1 ALIGNED WITH EU VALUES

All EC evaluations should be of high quality and in line with the principles spelled out in the EC's Better Regulation Guidelines (EC, 2021a, pp. 26–27):

- **Comprehensiveness.** Evaluations should, in general, cover seven evaluation criteria – the Organisation for Economic Co-operation and Development Development Assistance Committee's (OECD DAC) six criteria of relevance, coherence, effectiveness, efficiency, impact and sustainability and the EU-specific criterion of EU added value. Not all criteria are relevant to every evaluation, and other criteria may be added as appropriate.
- **Proportionality.** The scope and analysis of evaluations must be appropriate and well-suited to what is being evaluated (the evaluand), its maturity and the data available.
- **Evidence-based.** Evaluations are based on the best available evidence (factual or opinion based) drawn from a diverse range of methods and sources (triangulation). Any limitations to the evidence used and the methodology applied – particularly in terms of ability to support the conclusions – must be clearly explained.
- **Transparent judgement.** Evaluators must make judgements based on the evidence and analysis available. These judgements should be as specific as possible, and the judgement criteria clearly identified during the design of the evaluation.
- **Independence and objectivity.** Robust and reliable results can be delivered only through independent and objective evaluation. The evaluation team must be free of bias and conflicts of interest, be able to carry out their work without pressure, given full access to needed information, and have full autonomy in reporting their findings.

The EU Evaluation Policy cites two complementary purposes of evaluation: **learning** – including about what works and what does not and under what conditions, facilitating evidence-based decision-making, and sharing experiences and good practices within the EU and with other partners – and **accountability** to stakeholders – including in the interests of transparency and by explaining the difference between what was planned and what was achieved. Either or both of these may be the focus of any evaluation, depending on user needs (EEAS and EC, 2014).

The two main purposes of evaluation – learning and accountability – lead to better and more timely decision-making, and enhance institutional memory on what works and what does not in different situations (see Figure 1.1).

FIGURE 1.1 Purposes of evaluation

NOTE: *Remember that evaluation is a learning process supporting decision-making at all levels of accountability; it should never be approached as a box-ticking exercise.*

1.1.2 PART OF A SYSTEM

Evaluation is part of the EC's overall monitoring and evaluation (M&E) system. Although serving different purposes, monitoring and evaluation are complementary assessments involving data collection, performance assessment, reporting and learning.

- **Monitoring** focuses on **what** has happened. It is a continuous and organised process of systematic data collection (or access) throughout the life of an initiative to oversee its progress. It generates information that feeds into future evaluation and impact assessments and provides a solid evidence base for policymaking. Monitoring data are validated by results-oriented monitoring (ROM) reviews (see Box 1.1) and inform both individual evaluations and the EU's overall Global Europe Results Framework (GERF).
- **Evaluation** identifies and explains not only what changes – intended or unintended – have occurred, but **how** and **why** they have occurred and what **learning** can be derived from that. The Better Regulation Toolbox explains that:

> Evaluation goes beyond an assessment of *what* has happened; it considers *why* something has occurred and if possible, *how much* has changed as a consequence. It thus aims (where possible) to draw conclusions about the causal effects of the EU intervention on the desired outcomes. It should also look at the wider perspective, seeking to identify (and learn from) any unintended/unexpected effects... (EC, 2023, p. 378).

Table 1.1 delineates these differences between monitoring and evaluation.

NOTE: *Neither monitoring nor evaluation is to be confused with an audit. An audit looks at the integrity of processes, procedures and compliance.*

1.1.3 LINKED TO A CYCLE

Evaluation is **integral to the intervention cycle and its management**, providing key inputs and information at all phases of the cycle (Figure 1.2). By providing evidence of what works and what does not (and under what circumstances), evaluation helps improve engagement with partners; enhances the impact of EU development cooperation; and is critical to better programming of subsequent interventions, political dialogue and visibility of results. The benefits of the knowledge generated through evaluation extend beyond the commissioning unit or delegation to be a source of institutional learning and living memory now and in the future. A good evaluation supports **better decision-making for better outcomes**.

1.1.4 AIMED AT USER NEEDS

The **primary intended users** of an evaluation are those individuals or groups that have a particular

BOX 1.1 Results-oriented monitoring

Established in 2001, ROM aims to enhance the quality of the EC's international partnership operations across the world. Although coordinated by EC services, ROM is **external** since it is provided through independent consultants (ROM contractors). It is **results-oriented** because it focuses on results and achievements, reflecting the growing attention of the EU and its Member States on the effectiveness and performance of their interventions. ROM provides snapshots rather than the in-depth analysis provided by evaluation. It is primarily used by project managers and implementing partners for course correction and adaptation.

ROM combines a methodology and set of adapted services aimed at strengthening the accountability and results-based management capacities of the EU and its partners. It provides:

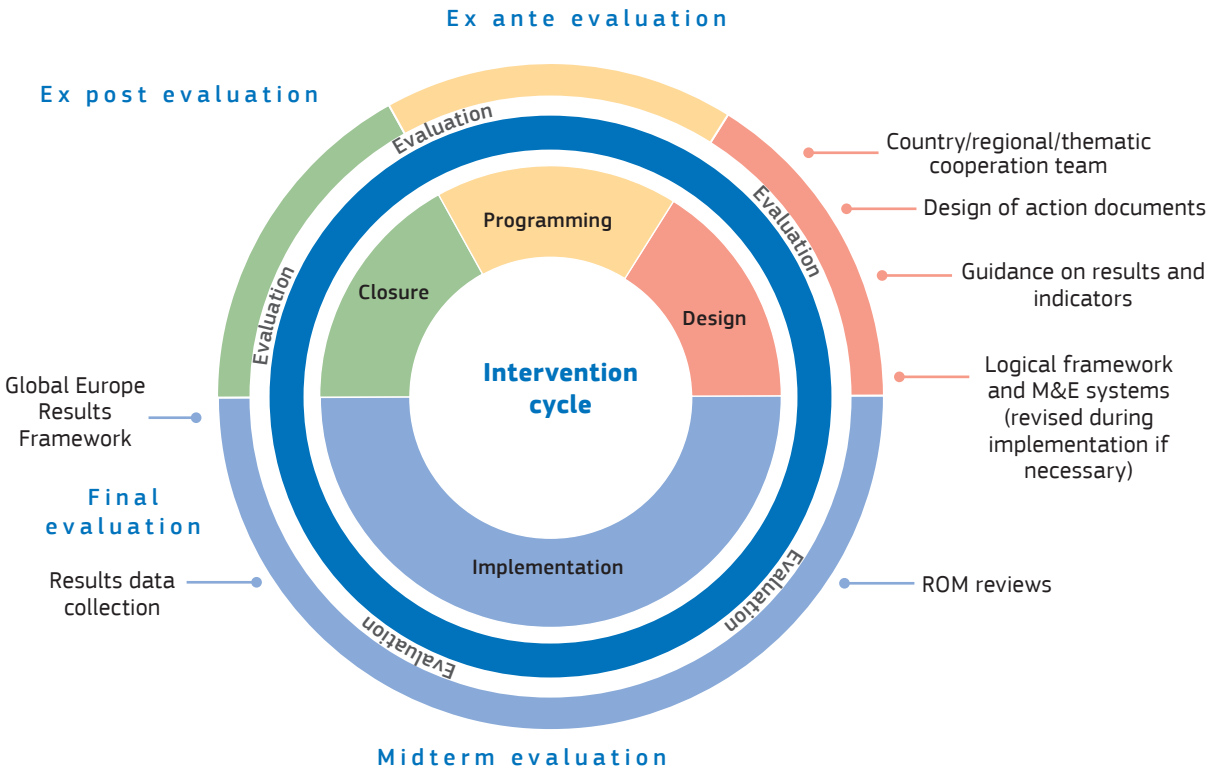
- support in the design of logical framework matrixes (logframes), M&E systems and reporting;
- support to results data collection for internal monitoring and reporting;
- ROM reviews at the intervention level to help steer implementation, keep progress on track and enable learning from difficulties.

For more information on ROM, click *here*.

TABLE 1.1 Monitoring, ROM and evaluation

	Monitoring	ROM	Evaluation
What	Daily management activity (piloting the operation)	Ad hoc review carried out according to a standard methodology	Analysis for in-depth assessment
Who	Internal management responsibility – all levels (EC and implementing partner)	Always incorporates external inputs/resources (objectivity)	Usually incorporates external inputs/resources (objectivity)
When	Ongoing	Periodic – on demand or if intervention is facing problems	Ex ante, periodic (midterm, final), ex post
Why	Check progress, take remedial action, update plans	Check progress, take remedial action, provide input to follow-up actions	Learn broad lessons applicable to other interventions, policy review etc.
Focus	Inputs, activities, outputs, outcomes	Rationale, relevance, outcomes, sustainability, coherence, EU added value	Rationale, relevance, outcomes, impact, sustainability, coherence, EU added value and other criteria as relevant

FIGURE 1.2 The M&E system in the intervention cycle



interest or stake in the evaluation results and that will be making decisions based on its conclusions and recommendations. Evaluation users are generally policymakers and intervention designers, managers and operators in charge of implementation, partners, institutions that have provided financing and to which accountability is required, public authorities, civil society organisations and field practitioners.

Evaluation users look to evaluations to:

- **contribute to the design** of policy, programming documents or interventions;
- **analyse the added value** of a specific strategy, policy, instrument, modality or intervention;
- **analyse the value** of a strategy or cooperation with a given country or group of countries;
- **inform resource allocation**;
- **improve the quality of implementation** of a strategy, policy, instrument, modality or intervention – whether during the current cycle of implementation or in subsequent cycles;

  NOTE: *Evaluations that aim to improve quality of implementation are called* formative evaluations.

- **report on achievements** of and lessons from a strategy, policy, instrument, modality or intervention (learning and accountability);

  NOTE: *Evaluations that report on achievements are called* summative evaluations.

- **inform future** strategies, policies, instruments, modalities or interventions, including through scale-up, follow-up interventions and expansion of successful pilots.

Evaluations also enable interventions to **account for the use of financial and non-monetary resources**. At times, evaluations may **justify adaptations** during implementation and provide the rationale for such adaptations.

Box 1.2 summarises evaluation uses and users.

Beyond generating findings, conclusions and recommendations, evaluation processes in and of themselves can have positive effects on individuals, organisations and networks. When done well, the processes used in an evaluation – such as engaging with key stakeholders – can generate benefits such as increased trust and ownership of the evaluation results.

Evaluations should be responsive to gender equality and the empowerment of women and should assess whether interventions have been guided by international normative frameworks for gender equality; have analysed and assessed the structures that contribute to inequalities experienced by women, men, girls and boys and to those experiencing multiple forms of exclusion; have maximised participation and inclusion; and have sought to empower rights holders and duty bearers.

SEE: Box 1.3.

The process of doing an evaluation can:

- clearly signal to all stakeholders what is valued and what the priorities are so they can focus their efforts on what is most important (and hence improve performance based on evidence);
- strengthen participants' capacity to sustain activities and/or impacts into the future by increasing stakeholders' skills and knowledge and their connections to each other.

BOX 1.2 Evaluation uses, by user

Decision makers, **policymakers** and **intervention designers** use the evaluation to reform or renew the intervention, confirm or change strategic orientations, or (re)allocate resources (financial, human and others). They appreciate clear, simple and operational recommendations based on credible factual elements.

Managers and **operators** in charge of implementation use evaluation findings to adjust management, coordination and/or their interactions with beneficiaries and target groups. They expect detailed information and are ready to interpret technical and complex messages.

The **partners** and **institutions** that funded the intervention expect to receive accounts – that is, a conclusive overall assessment of the intervention.

Public authorities conducting related or similar interventions may transfer and adapt the lessons learned from the evaluation. The same applies to **field practitioners** and **expert networks** in the concerned sector.

Finally, the evaluation may be used by **civil society actors**, especially those representing the interests of the targeted groups, to lobby for change, replicate successful experiences, and enhance engagement and ownership by target groups.

BOX 1.3 Gender-responsive evaluation

Gender equality is a core value of the EU, emphasised and promoted in foundational documents from its establishment to the present day (see e.g. the European Pillar of Social Rights). Gender equality is a universally recognised human right, as well as an imperative to well-being, economic growth, prosperity, good governance, peace and security. As stated in the latest EU Gender Action Plan (GAP III for 2021–2027; EC, 2020a), 'All people, in all their diversity, should be free to live their chosen life, thrive socially and economically, participate and take a lead as equals'. GAP III, and its encompassing EU Gender Equality Strategy 2020–2025 (EC, 2020b) and the complementary LGBTIQ Equality Strategy 2020–2025, calls for a gender-equal world.

The objective of gender-responsive evaluation is to guide management and decision-making processes by providing information on the different ways in which EU external action is affecting women and girls on the one hand and men and boys on the other, thereby contributing to the achievement of gender equality commitments. It is applicable to all types of EU external action and development cooperation interventions and programming, not just gender-specific ones.

Gender equality and women's and girls' empowerment are long-term endeavours. Progress towards gender equality and women's empowerment is rarely straightforward and often accompanied by setbacks and new constraints. Gender equality and women's empowerment comprise many dimensions – voice/participation/agency (distribution of power), access to/control over resources/opportunities, and shifts in formal (legislation, policy etc.) and informal institutions (values and attitudes etc.) and social protection systems. Advancement in these dimensions is interlinked. Progress in one dimension may be hampered if efforts in another dimension are constrained. Important aspects of each of these dimensions are not easily measured. Evaluations seeking to measure progress towards gender equality and women's empowerment need to adopt a mix of quantitative and qualitative methods and participative approaches appropriate to measuring and evaluating social change (EC, 2024). A gender-responsive evaluation should include three elements:

- an assessment of the contribution that an intervention or policy has made towards the ultimate goal of gender equality;
- an assessment of the extent to which an intervention or policy has pursued gender mainstreaming to ensure that the concerns, experiences, practical needs and strategic interests of women and men, and girls and boys are equally addressed;
- an assessment of the extent to which an intervention or policy has been guided by international standards for gender equality and has analysed and addressed structures that contribute to inequalities, maximised participation and inclusion, and sought to empower rights holders and duty bearers.

For more information, see Evaluation with Gender as a Cross-cutting Dimension (EC, 2024).

1.1.5 DG INTPA AND FPI ROLE IN EVALUATION

The steering, supporting and coordinating functions of all evaluation activities are currently carried out by the Quality and Results, Evaluation, Knowledge Management unit (D4) in DG INTPA and the Budget, Finance, Relations with other Institutions unit (4) in FPI. With specific reference to evaluation functions, these units:

- support and coordinate the evaluation of interventions directly managed by the operational services;
- plan and manage strategic evaluations;
- support dissemination and follow-up on the results of evidence-based evaluations in policy and practice;
- develop methodologies, tools and staff capacity in evaluation, coordinating and working in partnership with internal and external stakeholders, including EU Member States' evaluation services and those of other development partners.

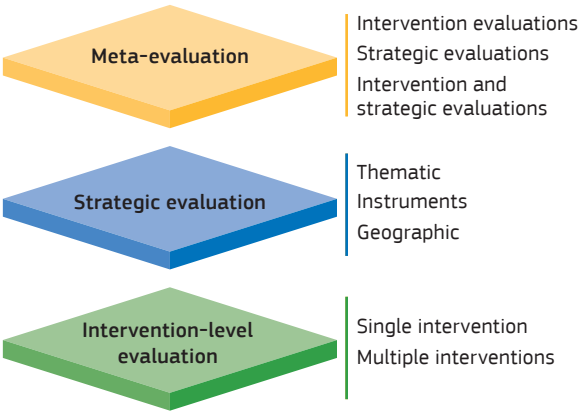
1.2 Evaluation types and timing

Evaluations can be done for different purposes and at different times to assess what works/worked and why. This section breaks down the different types of evaluation:

- **Intervention-level evaluations** – previously known as project and/or programme evaluations – analyse a specific intervention, or group of interventions, in one or multiple countries.
- **Strategic evaluations** look at the combination of the EU's external spending and non-spending actions to review EU strategies, policies, instruments or modalities, generally over a significant period of time.
- **Meta-evaluations** are used to provide a systematic analysis of existing evaluations (intervention level and/or strategic) to bring together core learning on similar topics.

Figure 1.3 summarises the different levels of evaluation.

FIGURE 1.3 Levels of DG INTPA evaluations



1.2.1 INTERVENTION-LEVEL EVALUATIONS

Intervention-level evaluations analyse the results of a specific intervention, or a group of logically interlinked interventions, within the frame of a wider scope of collaboration in a country or region. These evaluations are an integral part of intervention cycle management as they help enhance the programming, design, implementation, performance and achievement of results of EU interventions.

- **Evaluation of single interventions.** An evaluation of this type covers only one intervention, whether a single small grant, a component of a larger initiative, or a large-scale multimillion-euro effort spanning multiple years over several countries.
- **Evaluation of multiple interventions.** These evaluations cover several interventions included in the same or successive programming cycles. The grouped interventions must be clearly interlinked in a logical and unambiguous way – for example, their expected contribution to a common (or very similar) **overall objective** through the achievement of a set of consistent **outcomes**. The grouping of loosely interlinked interventions (or of interventions that are not logically interconnected) under a single evaluation is discouraged, as it disperses the focus of evaluators in a series of parallel and inconsistent analyses.

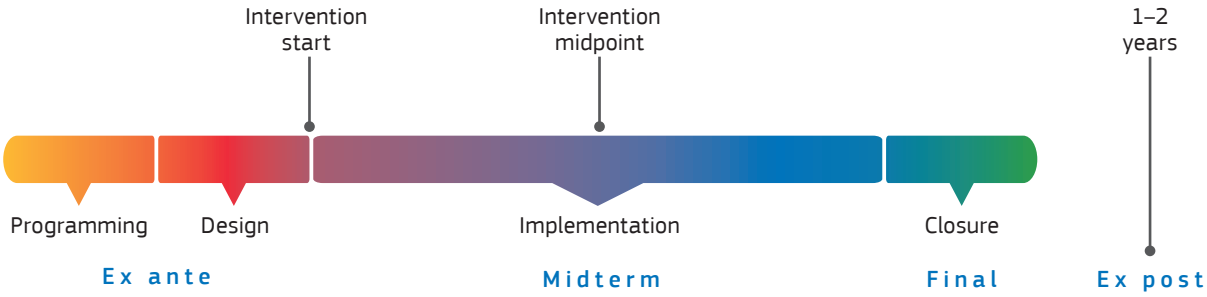
Intervention-level evaluations can be conducted before an intervention starts (ex ante evaluation), and/or at the midpoint (midterm evaluation) or conclusion (final evaluation) of the intervention (i.e. six months before/after the intervention's completion date), or following implementation (ex post evaluation) (i.e. at least one year after the intervention's completion date), as illustrated in Figure 1.4.

1.2.2 STRATEGIC EVALUATIONS

A strategic evaluation has a wider scope of analysis than an intervention-level evaluation and looks deeper into the strategic dimensions of thematic areas, instruments or overall EU cooperation and partnerships in a defined geographic area. More precisely, strategic evaluations analyse EU strategies from conception to implementation at any or all of several levels – country, region, sector or financing instrument – over an extended period of time (often 7–10 years).

NOTE: *Under special circumstances, a strategic evaluation can be launched to respond to a* ***pressing knowledge objective*** *under a tight deadline and*

FIGURE 1.4 Types of evaluation by intervention stage



with a narrow focus of analysis – for example, the Fast-Track Assessment of the EU Initial Response to the COVID-19 Crisis in Partner Countries and Regions (2020).

Sometimes, strategic evaluations can **combine different types of assessment**, such as geographic and thematic aspects.

EXAMPLE: Evaluation of the EU Regional Development Cooperation with Latin America (2009-2017).

A few further general points on strategic evaluations follow:

- Since these evaluations combine interventions and/or strategies at different degrees of implementation, they can be considered midterm or, when the work in a given sector/theme is in the process of phasing out, final.
- DG INTPA strategic evaluations have been conducted since 2007; those conducted since 2014 are available online on the DG INTPA Strategic Evaluation Reports web page.

SEE: *EC Evaluation in practice web page for more information about strategic evaluations at DG INTPA.*

Thematic evaluation

The 'theme' covered by a thematic evaluation can either be a sector of intervention, a policy area covering several sectors or a cross-cutting issue such as the Evaluation of the Green Deal planned for 2025.

NOTE: ***Cross-cutting issues*** *are those relevant to all aspects of an intervention (planning, design, implementation etc.) and/or more widely to the overall rationale of EU development cooperation policy.*

A thematic evaluation can analyse the results of one or several EU cooperation areas under a specified **sector of intervention** – energy, migration, infrastructure, governance etc.; it can cover multiple countries or regions or be global in scope.

A thematic evaluation can also look at a specific set of interventions from the point of view of a **cross-cutting issue** – generally, a policy objective having a medium- to long-term perspective such as gender equality, environmental protection, equity, inclusion or human rights. These thematic evaluations are also conducted with a wide geographic scope.

EXAMPLE: Evaluation of the EU's External Action Support in the Area of Gender Equality and Women's and Girls' Empowerment (2010–2018).

Instrument, mechanism and modality evaluation

This type of strategic evaluation aims to understand the results of using specific **financing instruments** established by EU regulations (such as the Neighbourhood, Development and International Cooperation Instrument – Global Europe and the European Fund for Sustainable Development Plus), **financial mechanisms** (such as blending and budgetary guarantees) and **modalities** (such as budget support; see Box 1.4) across regions or worldwide. The objective of this type of evaluation is to understand the value of a specific instrument, mechanism or modality and its added value in comparison with other forms of partnership and aid delivery.

BOX 1.4 Budget support evaluations

Budget support is a means of delivering effective aid/assistance and durable results in support of EU partners' reform efforts and the Sustainable Development Goals. In 2020, EU budget support disbursements amounted to EUR 3 billion, accounting for 24 per cent of total EU external assistance.

What differentiates budget support from more traditional interventions is that funds are channelled directly through the partner country's treasury; hence it provides key opportunities to enhance a partner country's capacity to develop and implement its own policies, whereas traditional interventions provide a combination of pre-defined goods and services.

Evaluation of budget support implies some modifications to the standard evaluation methodology. For example, its intervention logic recognises that actual changes in terms of policy outcomes are the responsibility of governments and civil societies in the context of given 'opportunity frameworks'. External assistance can contribute to this change process by creating specific additional opportunities, which can be appropriated by the targeted partners and adapted to their own context in different ways. An evaluation must therefore be able to assess the extent to which such opportunities have been relevant and have materialised, and how they have been appropriated and adapted by the partner governments in order to achieve the results laid out in their strategies and targeted by budget support.

For more information on budget support evaluations, see the Annex to this handbook.

EXAMPLES: Evaluation of EU Budget Support and Blending in the Kyrgyz Republic (2010–2019); Mid-term Evaluation of the European Union Emergency Trust Fund for Stability and Addressing Root Causes of Irregular Migration and Displaced Persons in Africa 2015–2019.

Geographic evaluation

Geographic evaluations look at the entire EU's external spending and non-spending actions at the country or regional level. Their objective is to analyse the results of EU cooperation or partnership with a country or group of countries over a defined period of time, generally covering an entire financing cycle or multiple financing cycles.

EXAMPLE: External Evaluation of the European Union's Cooperation with Myanmar (2012-2017).

1.2.3 META-EVALUATIONS

A meta-evaluation is a systematic analysis of evaluations conducted to bring together core learning on similar topics (EEAS and EC, 2014). It can focus on either intervention or strategic evaluations or a combination of both.

Meta-evaluations synthesise **recurring findings, conclusions and recommendations from different evaluation reports** in order to provide inputs into strategic planning and institutional learning. The main types (and timing) of meta-evaluation are proactive, retroactive and concurrent.

- **Proactive meta-evaluations** are conducted before embarking on a full-scale evaluation to provide an overview of what is already known; this information should help in assessing the evaluability of a specific policy or instrument and define the subject and scope of the upcoming evaluation.

  EXAMPLE: Evaluation of the Civil Society Organisations and Local Authorities Thematic Programme (2014–2019).

  SEE: Subsection 3.1.1 *for information on conducting an evaluability assessment.*

- **Retroactive meta-evaluations** focus on a series of previously conducted evaluations and aim at aggregating the results from these evaluations as lessons learned and to feed into planning.

  EXAMPLE: Synthesis of Budget Support Evaluations: Analysis of the Findings, Conclusions and Recommendations of Seven Country Evaluations of Budget Support.

- **Concurrent meta-evaluations** are conducted during another strategic or intervention evaluation process. They constitute an integral part of the methodology used and aim at gathering secondary data, namely evidence from previous evaluations.

SEE: Table 1.2 *for a summation of evaluation purpose, timing, type and users. For more detail on budget support, see the* Annex.

1.3 Who conducts evaluations?

Many people and entities conduct evaluations, including relevant EU services (including DG INTPA and FPI), the EU delegations, partners, government authorities and civil society organisations. The various types of evaluations described in Section 1.2 can be conducted either by the EC alone, generally through framework contracts, jointly with one or more partners, or by a partner (third-party evaluation).

1.3.1 EVALUATIONS MANAGED BY DG INTPA AND FPI

Responsibility for EC-managed evaluations lies with the EU delegations/regional teams (FPI) or headquarters units; they plan, launch, contract for and manage these evaluations. Evaluations of **interventions** are generally the responsibility of the delegation/unit in charge of the intervention; for **strategic** evaluations, the responsibility generally lies with DG INTPA.D.4/FPI.4. These evaluations are contracted for in OPSYS by launching a request for services under a framework contract.

NOTE: *Framework contracts are a contractual tool that allows to mobilize rapidly (compared to a standard tender procedure) through specific contracts the expertise required to assist*

TABLE 1.2 Matrix of evaluation purposes, timing, types and users

Purpose	Timing and type	User
Improve the quality of a current intervention (or policy, instrument, country-level cooperation etc.)	• Intervention-level midterm • Strategic	• Operational managers, EU delegations and central units managing evaluated intervention • Implementing partners • Steering committees of evaluated intervention (including representatives from partner government) • Senior decision makers • Other related bilateral/multilateral agencies
Accountability – report on the achievements of an intervention (or policy, instrument, country-level cooperation etc.)	• Intervention-level final • Intervention-level ex post • Strategic	• EU policymakers and senior decision makers • Partner governments • Funders • European Parliament and European Council • EU Member States • Wider public • Final beneficiaries and communities
Inform **resource allocation** for future interventions (or policy, instrument, country-level cooperation etc.)	• Intervention-level midterm • Intervention-level final	• Operational managers, EU delegations and central units managing evaluated intervention • Implementing partners • Steering committees of evaluated intervention (including representatives from partner government) • Senior decision makers • Other related bilateral/multilateral agencies
Inform **future** interventions, including scaling-up, follow-up interventions and translation of successful pilots	• Ex post • Strategic • Meta-evaluation • Final	• Bilateral/multilateral agencies • DG INTPA, FPI, EEAS and other EU DGs and services • Partner countries' governments and regional authorities • EU delegation counterparts (line ministries in partner countries) • Implementing partners • Research community • Wider public

the European Commission departments in implementing/evaluating their policies. See the Framework Contract web page.

For strategic evaluations, DG INTPA.D.4 – in consultation with all relevant headquarters units, delegations and the European External Action Service – prepares a three-year rolling **Evaluation Work Programme** (EWP), which is approved by DG INTPA management. The EWP establishes the list of countries, regions, instruments, themes, modalities and policies that will be subject to strategic evaluation in the upcoming three-year period.

NOTE: *In DG INTPA, evaluation of **budget support** can be either strategic or at the intervention level. The former are launched and managed by DG INTPA.D.4. Evaluations of one or more budget support interventions are launched and managed by the delegations/units.*

1.3.2 JOINT EVALUATIONS

Joint evaluations of jointly funded interventions can be carried out, in whole or in part, by the EU and one or more partner organisations and/or EU Member States. The scope of the interaction and the nature and mechanism of exchange vary; for example, joint evaluations could take the form of:

- **joint data collection and/or exchange of assessments** with external actors, where each partner conducts its own analyses and prepares a separate report;
- **collaborative evaluation**, where each partner is mutually and equally responsible for evaluation design, implementation and the development of joint recommendations.

Joint evaluations, while potentially more time-consuming due to the need for coordinating among various stakeholders, are often highly productive. These evaluations can help overcome attribution problems in assessing the effectiveness of interventions and strategies, the complementarity of efforts supported by different partners, the quality of aid coordination etc.

NOTE: *The decision to conduct a joint evaluation is up to the delegation/unit. See Section 1.4 for more information.*

1.3.3 THIRD-PARTY EVALUATIONS

Third-party, or partner-led, evaluations are contracted to or conducted by different entities depending on the intervention's management (see Box 1.5):

- interventions under **indirect management**, such as pillar-assessed organisations – entrusted entities;

  NOTE: *EC partner organisations must pass pillar (institutional compliance) assessments as a prerequisite to indirect management cooperation. See International Partnerships Audit and Control web page.*

- interventions under **direct management** financed with EU funds, such as grants to civil society

BOX 1.5 Different management modes: direct and indirect management

In **direct management** mode, the EC is directly responsible for all steps in an intervention's implementation (i.e. launching the calls for proposals, evaluating submitted proposals, signing grant agreements, monitoring project implementation, and assessing the results and making payments).

The majority of the EU budget allocated to international development is implemented under **indirect management**. Under this management mode, the Commission delegates budget execution tasks to different types of implementing partners, including international organisations such as United Nations entities, the World Bank, the International Monetary Fund, the European Investment Bank and the European Investment Fund; and Member States' development agencies such as the Spanish Agency for International Development Cooperation (AECID) and the French Development Agency (AFD). These partners are often referred to as '**pillar assessed**'. As stipulated in the EU Financial Regulations, the EC can decide to entrust implementing partners with budget implementation tasks. For organisations to become eligible, they must meet certain conditions and pass an assessment of their main operations by an independent audit or to determine compliance with EC requirements for indirect management.

organisations or non-governmental organisations – intervention implementers.

The present legal framework specifies that evaluations of indirectly managed interventions implemented by entrusted entities be carried out by those agencies according to their rules, while the evaluation of interventions under direct management such as grants should be carried out as defined in the corresponding contract (e.g. grant agreement). The EC reserves the right to conduct the evaluation of these interventions itself.

1.4 Putting it together: evaluation planning

An evaluation can be planned and carried out at any time of the year in line with the needs of the delegations and headquarters units in the case of intervention-level evaluations, and on the basis of the three-year rolling EWP for strategic evaluations. This section provides information on how to make choices about exactly what and when to evaluate.

1.4.1 WHAT TO EVALUATE

Each year, the EU delegations and headquarters units plan and/or revise their projected evaluations according to their priorities, needs and resource availability. By regulation, EU delegations and headquarters units are mandated to evaluate those interventions that represent a significant investment of funds.

> **NOTE:** *'Programmes and activities that entail significant spending shall be subject to ex-ante and retrospective evaluations ("evaluation"), which shall be proportionate to the objectives and expenditure' (Article 29,* Commission Delegated Regulation (EU) 2019/715 *on the framework financial regulation).*

Beyond this requirement, and aside from those interventions whose evaluation is contractually mandated, the decision on what to evaluate is completely at the discretion of each delegation/unit. A best practice is to select a **well-thought-out sample of interventions** for evaluation; Box 1.6 presents some rationales to help in prioritising strategies, policies, instruments, modalities or interventions for evaluation.

As mentioned, there is no limit to the **number of interventions** that can be covered by a single evaluation – to the extent that this remains feasible and there is a sound justification to evaluate multiple logically interconnected/interrelated interventions together.

> **NOTE:** *Evaluability assessment can be used to support decisions about what to evaluate and when. See* Subsection 3.1.1.

But how is the selection to be made? Basic, **first-line considerations** are to ensure that:

- evaluations are conducted at the **right time** to be useful as learning and decision-making tools;
- a significant portion of the portfolio is evaluated for **accountability** and transparency purposes, considering the regulation cited above;
- resources to be devoted to managing the evaluations are **efficiently allocated** within each unit and delegation.

> **NOTE:** *Be realistic. Be aware of available resources when deciding which evaluations can and will be implemented.*

1.4.2 WHEN TO EVALUATE

To make the most of an evaluation, determine its best **timing** with regard to the intervention cycle. See Figure 1.5 for guidance in making this determination.

BOX 1.6 Issues to consider in choosing what to evaluate

TIMELINESS

- **Active demand.** Consider prioritising evaluations where there is active demand (e.g. from stakeholders or EU citizens) for evidence to inform decisions and where evaluations can be timely in terms of decision-making.
- **Linking programming and design.** Consider prioritising evaluation of interventions providing useful evidence for better design of upcoming interventions or policy strategies.
- **Contextual changes.** Consider prioritising those interventions where important changes in the context call into question continuation of the intervention as initially planned.

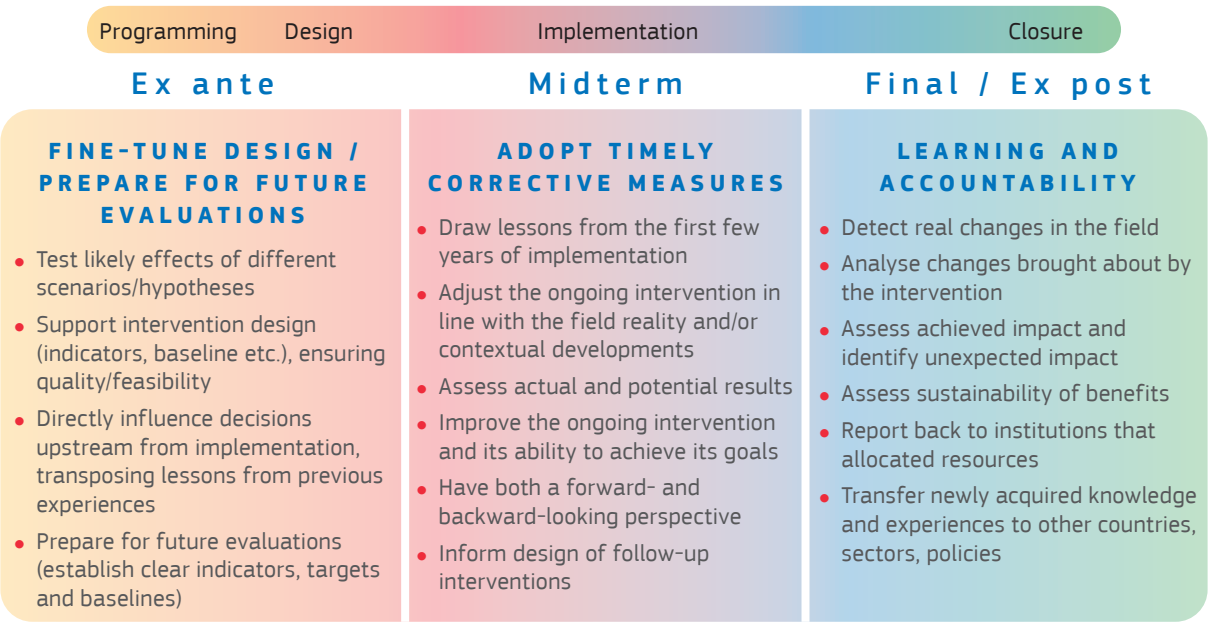
ACCOUNTABILITY

- **Significant spending.** Consider prioritising those interventions that entail significant spending, as this responds to a general duty of accountability for the use of public funds.
- **Risk management.** To better manage risk, consider prioritising interventions with significant investment (of money, human resources, time or community goodwill), or where there is seen to be high risk in terms of either failing to achieve intended impacts or producing significant negative unintended impacts (e.g. in contexts of fragility, conflict and violence or when dealing with sensitive issues such as sexual and gender-based violence).

EFFICIENCY

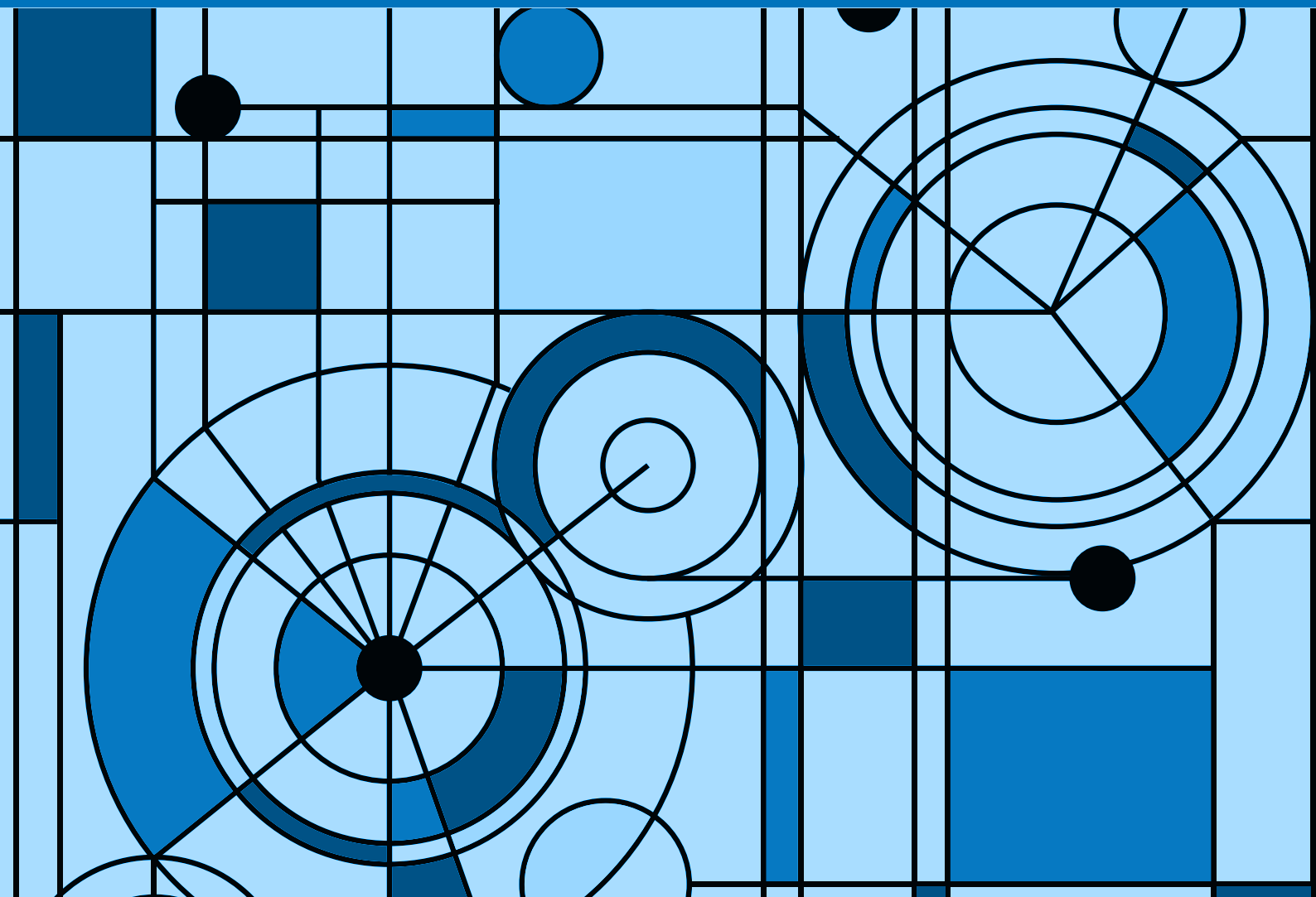
- **Best opportunities for learning.** Consider prioritising evaluations of interventions and topics that fit into the current or upcoming priorities of your delegation/unit, are innovative in their specific context or seemed particularly successful or unsuccessful in meeting their objectives.
- **Potential for replication and scale-up.** Consider prioritising opportunities to explore the replication or scale-up of an intervention in the same context or elsewhere.
- **Value for money: less is more.** Consider prioritising interventions where the cost of evaluation is reasonable, given the cost of the intervention itself.
- **Value for money: more is more.** Consider prioritising the opportunity to carry out thematic evaluations of several interventions in the same sector/subsector. Simultaneously evaluating multiple interventions that work in the same area, even if they are at different stages of implementation, can be highly meaningful from a strategic point of view and would contribute towards reducing the cost of evaluation and evaluation fatigue.

FIGURE 1.5 Determine evaluation timing



2

Managing an evaluation



What is this chapter about?

This chapter provides hands-on practical guidance for evaluation management across the entire process of an evaluation – its preparation, implementation and follow-up activities.

How will this help you in your work?

This chapter takes you through the different phases of an evaluation, detailing the tasks to be accomplished, the steps to be followed and the templates – mandatory or recommended – to be used.

For definitions of key terms used in this handbook, refer to the glossary.

SECTION 2.1

Evaluation phases and stakeholders

Chapter 2 explains how to manage an evaluation through all its phases – from the preparation to launch an evaluation to the dissemination of and follow-up on its results. It details the processes and outputs as well as the key stakeholders and their roles and responsibilities.

Step-by-step guidance and clear explanations are provided, along with links to relevant resources, templates, documents, and other references, and to the OPSYS Portal.

Although the chapter is organised chronologically through the six phases – preparatory, inception, interim, synthesis, dissemination and follow-up – the structuring of any evaluation is **adaptable**, depending on such factors as type of evaluation, scope, financial resources available and security conditions in the field.

While the chapter focuses on the **strategic and intervention-level evaluations** undertaken by European Union (EU) delegations and headquarters, the main steps and responsibilities presented here are **valid for other types of evaluations** as well (joint evaluations, evaluations carried out by partners etc.).

As essential **context** for the following sections of this chapter, this introductory section provides an overview of:

- the six **phases** of a typical evaluation;
- the four main categories of **stakeholders** in an evaluation.

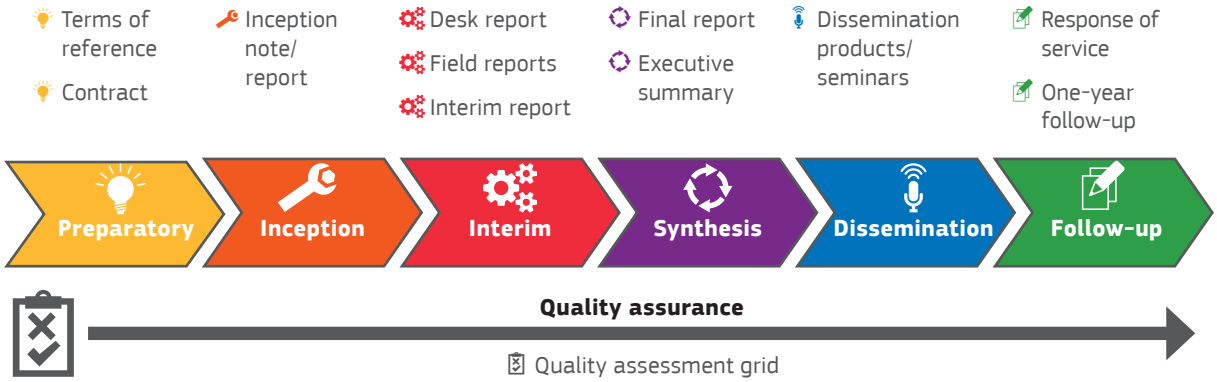
2.1.1 The six phases of an evaluation

The six standard phases of an evaluation are described below and illustrated, along with their associated main deliverables, in Figure 2.1.1.

- **Preparatory phase.** In this phase (covered in Section 2.2), the evaluation mandate is defined. The evaluation manager sets up the evaluation reference group, drafts the terms of reference (ToR) and carries out the contractual procedures to recruit the evaluation team.
- **Inception phase.** In this phase (covered in Section 2.3), the evaluation is structured and the key issues to be addressed are clarified. The evaluation team, guided by the evaluation manager and the reference group, develops the final list of evaluation questions, the evaluation matrix and the evaluation methodology.
- **Interim phase.** In this phase (covered in Section 2.4), data and information are gathered and analysed to respond to the evaluation questions. As agreed with the reference group, the evaluation team conducts desk and/or field activities:
  - **Desk activities** are aimed at analysing the relevant data, drafting preliminary answers to the evaluation questions and identifying hypotheses to be tested in subsequent phases; activities typically include reviewing documentation, interviewing key stakeholders and other initial data activities using various tools (e.g. surveys).
  - **Field activities** are devoted to conducting primary research and further data collection to validate/modify the hypotheses formulated during the desk activities.
- **Synthesis phase.** In this phase (covered in Section 2.5), major evaluation findings derived from the evaluation questions, conclusions and recommendations are captured in a draft final report. The evaluation team compiles findings evolving from responses to the evaluation questions and formulates conclusions and recommendations and any relevant lessons learned in a draft final report, including a stand-alone executive summary.
- **Dissemination phase.** In this phase (covered in Section 2.6), the evaluation manager ensures that the evaluation is useful for and used by relevant stakeholders by capturing findings in easy-to-access formats. Broad dissemination, with content and format tailored to the specific audiences, is highly recommended.
- **Follow-up phase.** In this phase (covered in Section 2.7), the evaluation manager ensures that mandatory follow-up actions are taken. This entails identifying and alerting the relevant service/stakeholder responsible for taking action on each of the evaluation's recommendations, and then following up a year later to determine the extent to which they acted on the tasks planned for each recommendation.

Quality assurance (covered in Section 2.8) of all evaluation methods, outputs and deliverables takes place throughout all phases of an evaluation. Quality is assured by different key stakeholders, including

FIGURE 2.1.1 Flowchart of the evaluation process: phases and outputs



the evaluation team, particularly the team leader; the quality controller designated by the evaluation contractor; and the reference group, particularly the evaluation manager. Notably, the evaluation manager and the reference group assess the quality of the draft final report and executive summary using the online quality assessment grid (QAG).

2.1.2 Evaluation stakeholders

There are four main categories of stakeholders involved in a typical evaluation – the evaluation manager, the reference group, the evaluation team and other stakeholders. A summary of their respective roles and responsibilities is provided below and illustrated in Figure 2.1.2.

EVALUATION MANAGER

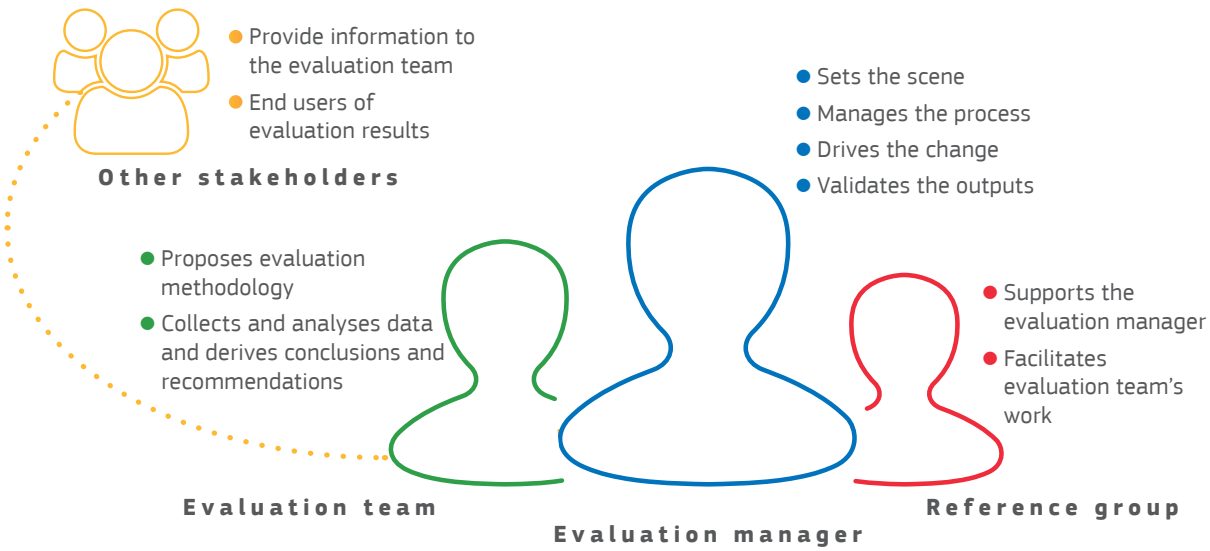
The evaluation manager is a member of the European Commission (EC) service commissioning the evaluation and is assigned by that service to manage the process on its behalf. The evaluation manager is responsible for **ensuring the quality and utility** of the evaluation – that is, ensuring that it meets the purposes set out in the ToR. The evaluation manager is also the person who has the final say in validating evaluation outputs and in giving the green light to move from one evaluation phase to the next. The evaluation manager's tasks include establishing the reference group; carrying out the contractual procedures – from drafting the terms of reference to selecting the evaluation contractor – recruiting the evaluation team; managing the evaluation process, including ensuring quality control; and driving the change resulting from the evaluation, including dissemination and follow-up.

REFERENCE GROUP

The reference group is presided over by the evaluation manager and provides assistance to the latter in guiding and supervising the evaluation. The reference group is composed of colleagues and stakeholders from members of EC services as well as representatives from partner countries and/or other organisations whenever possible. Reference group members should each be able to contribute a particular expertise and insight to **support and facilitate** the evaluation manager and team. The official membership of a reference group for intervention-level evaluations typically consists of between 3 and 6 participants; strategic evaluation reference groups can have between 5 and 10 official members. For evaluations of small interventions, a group may consist of as few as two members.

SEE: Subsection 2.2.1 *for more information about the role and significance of the reference group.*

FIGURE 2.1.2 Evaluation stakeholders



NOTE: *All evaluations, particularly complex ones, can benefit from review of deliverables and comment by colleagues and/or stakeholders from outside the reference group if and as needed.*

EVALUATION TEAM

The evaluation team is assembled by the selected contractor; it is responsible for developing the **evaluation methodology**, **data collection and analysis** as well as for the formulation of value judgements in response to the evaluation questions. The team writes and is responsible for the evaluation report. It submits its work regularly to the reference group and to the evaluation manager, and takes their comments into account. The contractor ensures ongoing support to the evaluation team as well as quality control over the team's deliverables.

OTHER STAKEHOLDERS

Other stakeholders are those individuals, groups and/or organisations not involved in the evaluation's management or implementation or part of the reference group that **have a direct or indirect interest** in the evaluated intervention and in the evaluation itself. They may or may not be affected by the intervention. Some may be a source of information for the evaluation; others may be end users of its findings. These other stakeholders are consulted and/or engaged with throughout the evaluation via workshops, focus group discussions, surveys, individual interviews etc.

Table 2.1.1 provides a summary of the functions, roles and responsibilities of all four categories of evaluation stakeholders by phase. This information is fleshed out in the remaining sections of this chapter.

TABLE 2.1.1 Evaluation responsibilities by phase and stakeholder

Evaluation manager	Reference group	Evaluation team (contractor)	Other stakeholders
PREPARATORY PHASE			
• Define the evaluation mandate • Plan evaluation and its timing to meet its purpose and use • Collect preliminary data • Invite, set up and chair the reference group • Formulate indicative evaluation questions and draw up the ToR • Carry out contractual procedures to recruit the evaluation team	• Provide input to the ToR • Aggregate and summarise views of stakeholders represented	Prepare and submit technical and financial proposals	
INCEPTION PHASE			
• Organise kick-off meeting • Support access to information and key stakeholders (e.g. provide evaluators with file of key documents noted in ToR) • Refine and finalise evaluation questions, (re)construct intervention logic or theory of change and finalise evaluation methodology • Organise inception meeting to present and discuss draft inception report • Approve reports/deliverables as per ToR (including revised methodology, evaluation questions and work plan)	• Participate in kick-off meeting to summarise expectations of stakeholders represented • Facilitate access to, and consultation by evaluation team of, all information sources and documentation on the evaluand • Agree on evaluation questions • Participate in the inception meeting and listen to/ discuss presentation of draft inception report • Comment on reports/ deliverables	• Listen to expectations of reference group at kick-off meeting • Propose/finalise evaluation questions, judgement criteria, indicators and data collection and analysis methods • Synthesise proposed methodology into evaluation matrix (including judgement criteria, evaluation indicators and data collection and analysis methods) • Develop work plan and refine distribution of responsibilities within the evaluation team • Produce inception report/note	

(continued)

TABLE 2.1.1 Evaluation responsibilities by phase and stakeholder (continued)

Evaluation manager	Reference group	Evaluation team (contractor)	Other stakeholders
INTERIM PHASE			
• Serve as the regular Interface with the evaluation team leader to support access to information and key stakeholders • Possibly organise planned debriefings and presentations • Approve reports/deliverables as per the ToR	• Act as interface between evaluation team and relevant stakeholders (facilitate contacts, interviews, access etc.) • Participate in any debriefing meetings and presentation of intermediate deliverables • Discuss and comment on desk/field/interim reports if requested by evaluation manager	• Carry out data collection and analysis • Cross-check data and information • Present any required interim debriefings/presentations • Produce any intermediate deliverables that may be required	• Be consulted through workshops, focus groups, individual interviews, surveys etc. • Participate in debriefing sessions if relevant
SYNTHESIS PHASE			
• Arrange discussions and debates on deliverables, conclusions and recommendations • Compile comments on the draft evaluation report for revision and ensure their integration into the final report • Approve reports/deliverables as per the ToR	• Discuss and comment on the various notes, reports and other products of the evaluation team • Aggregate and summarise views of stakeholders represented in the reference group commenting on the draft assessment report	• Produce judgements based on sound evidence and analysis • Formulate and articulate findings, conclusions and recommendations • Present the interim evaluation report • Include comments on the draft assessment report or justify their exclusion • Produce a final evaluation report with all required annexes and deliverables • Participate in various meetings and discussion seminars	Participate in debriefing sessions if relevant
DISSEMINATION PHASE			
• Plan dissemination activities for evaluation results and recommendations in collaboration with the evaluation team • Formally validate dissemination deliverables • Ensure good communication and dissemination of evaluation results and recommendations	Facilitate knowledge transfer through mobilisation of stakeholders around evaluation dissemination activities and deliverables	• Participate in various meetings and seminars for discussion/ dissemination of the evaluation • Produce dissemination outputs	Be targeted by the evaluation's dissemination activities and deliverables
FOLLOW-UP PHASE			
Follow up on recommendations	Play an active role in follow-up on evaluation findings, conclusions and recommendations		Be mobilised to implement some of the evaluation's recommendations

SECTION 2.2

Preparatory phase

The preparatory phase is when the **evaluation's mandate is defined**. This mandate provides details on the evaluation's temporal, geographic and legal scope; and in the case of strategic evaluations, on the country, region, sector or theme to be evaluated. During this phase and before preparing the terms of reference (ToR), the commissioning service clarifies in writing precisely what is to be evaluated and who the main intended users of the evaluation are.

The preparatory phase lays the foundation for carrying out and managing the evaluation. It is constructed around three main tasks:

- **Identifying the stakeholders.** The evaluation manager identifies relevant individuals or groups with an interest in the evaluation's subject and results. Some of these will be invited to serve as members of the evaluation reference group.
- **Defining the evaluation mandate.** With reference group support and some preliminary data collection, the evaluation manager defines the mandate of the evaluation, which is then spelled out in the evaluation ToR.
- **Managing the contractual preparations.** The evaluation manager launches the request for services, logs and tracks progress in OPSYS, chairs the evaluation of the submitted offers, awards and issues the contract, and executes all other steps needed to move the evaluation to the next phase.

2.2.1 Setting up the reference group

One of the evaluation manager's first tasks is to determine which stakeholders should serve on the evaluation's reference group. This group is an essential resource – throughout the life of the evaluation – for both the evaluation manager and the evaluation team, as its members will, individually and collectively, provide **liaison, expertise and perspective**.

SEE: How-to Guide on managing a reference group *on the Evaluation wiki for more information.*

ROLE AND PURPOSE

A reference group helps ensure access to information, accuracy of interpretations, and ownership of conclusions and recommendations. It acts as an **interface** between the evaluation manager and the evaluation team, and between the evaluation team and other stakeholders, opening doors and facilitating access to people and to relevant information sources and documentation. Notably (but not exhaustively), the reference group:

- discusses and comments on the ToR – including the proposed evaluation questions – drawn up by the evaluation manager, thereby contributing to the relevance and ownership of the evaluation;
- validates the proposed evaluation methodology put forward by the evaluation team including evaluation questions, judgement criteria and tools/methods for data collection (see Section 2.3);
- plays an important supportive role in quality assurance (see Section 2.8), discussing and providing feedback on notes and reports produced by the evaluation team as well as on the findings, conclusions and recommendations arising from the evaluation.

NOTE: *Also see the functions listed in* Table 2.1.1.

MAKING THE SELECTION

The evaluation manager decides who should serve on the reference group, making the decision with an eye to **selecting people who will add value** and contribute to the quality/usefulness of the evaluation.

Who to include. Typically, the reference group includes representatives from other European Commission (EC) services, the partner government (central and/or local level), other development cooperation partners, experts, non-state actors such as non-governmental organisations (NGOs) or civil society organisations, and other qualified participants. Members should also be drawn from among the evaluand's **stakeholders** – that is, implementing partners (but see Box 2.2.1), national partners, target groups and beneficiaries.

How many people to include. Experience shows that a relatively small reference group (typically 3–6 participants or as few as 2 for a small intervention; and 5–10 for strategic evaluations) is far more productive than a larger one. Membership should be manageable, with an emphasis on quality rather than quantity. **Diversity** should be a consideration to ensure the perspectives of different stakeholder groups are included in the steering of the evaluation. Such diversity of opinion broadens and enriches the scope and depth of the evaluation and ensures that as many voices as possible are heard.

BOX 2.2.1 Should implementing partners serve on the reference group?

The inclusion of implementing partners/agents in the reference group is at the discretion of the evaluation manager, who will need to weigh the **benefits to be gained in terms of ownership** of evaluation results and recommendations if they are part of the reference group against the **potential impact on independence** if they are likely to try to steer the evaluation in a particular direction.

In certain cases, having implementing partners on the reference group could potentially create a difficult group dynamic, as they have a legitimate interest in showing that 'their' intervention is performing well – which might have a negative impact on attempts at constructive and open debate around the successes and failures of the evaluand. One solution is for **implementing partners to join some reference group meetings without voting rights** and to absent themselves when requested by the chair. The feasibility of this option should be assessed on a case-by-case basis.

2.2.2 Defining the evaluation mandate

A major task facing the evaluation manager and the reference group is to **explain the purpose of the evaluation** in sufficient detail so that a suitably qualified evaluation team can be recruited. This information is documented in the ToR (see Subsection 2.2.4).

> **NOTE:** *Evaluations are undertaken by independent evaluators who are contracted via a framework contract; for more information, see Framework Contracts (SEA 2023). For each planned evaluation, the ToR is sent to framework contractors through OPSYS in the form of a request for services.*

All of the evaluation mandate elements set out in the ToR will be revisited, honed and likely revised in the inception phase (see e.g. Subsection 2.3.2 and Subsection 2.3.3) when the full evaluation team is in place. Nonetheless, the ToR is the foundation of the evaluation and the **first means of communicating with the prospective contractor/evaluation team**. It is thus critical that the ToR be as clear and precise as possible to allow the participating companies to develop good-quality, relevant offers – and so all involved have a common baseline for proceeding.

RESOURCES

The evaluation manager should start by reading all available basic documents about the evaluand, including:

- financing agreements;
- programming documents (e.g. project fiches, action documents and any modifications to these);
- design documents (intervention logic, logical framework matrix and/or theory of change);
- internal and external progress and monitoring reports, including results-oriented monitoring (ROM) reviews, any ex ante evaluations;
- any other relevant evaluations and/or research studies carried out by civil society, government, other donors – especially European Union (EU) Member States – and/or the private sector;
- EU documents setting out the policy framework in which the intervention takes place (EU development and external relations policy, EU foreign policy, country strategy paper);
- government strategy (e.g. poverty reduction strategy paper).

> **NOTE:** *This reading and research informs completion of the ToR background section (see Subsection 2.2.4).*

The evaluation manager should also have **informal conversations** with a few key stakeholders/informants to gain better insights into the evaluand.

OTHER INPUTS

A timeline for carrying out the evaluation is set by the evaluation manager, in line with institutional requirements and the objectives of the evaluation.

This schedule should allow sufficient time for quality review and the inevitable iterative revision process that is part of all evaluations. Evaluation managers can thus accommodate unforeseen situations such as poor-quality deliverables without feeling pressured to move to the next phase before they are resolved. This extra time can officially be designated as a buffer, or can be built into the different evaluation phases.

DETERMINING OBJECTIVES AND SCOPE

Based on this preliminary research and the above inputs, the evaluation manager and the reference group should meet to discuss and establish the evaluation's overall objectives and scope (Figure 2.2.1).

When these elements are in clear focus, the evaluation manager is ready to begin drafting the ToR, in consultation with the reference group (Figure 2.2.2). The evaluation manager then finalises the document and launches the request for services (see Subsection 2.2.5).

FIGURE 2.2.1 Defining the evaluation objectives and scope

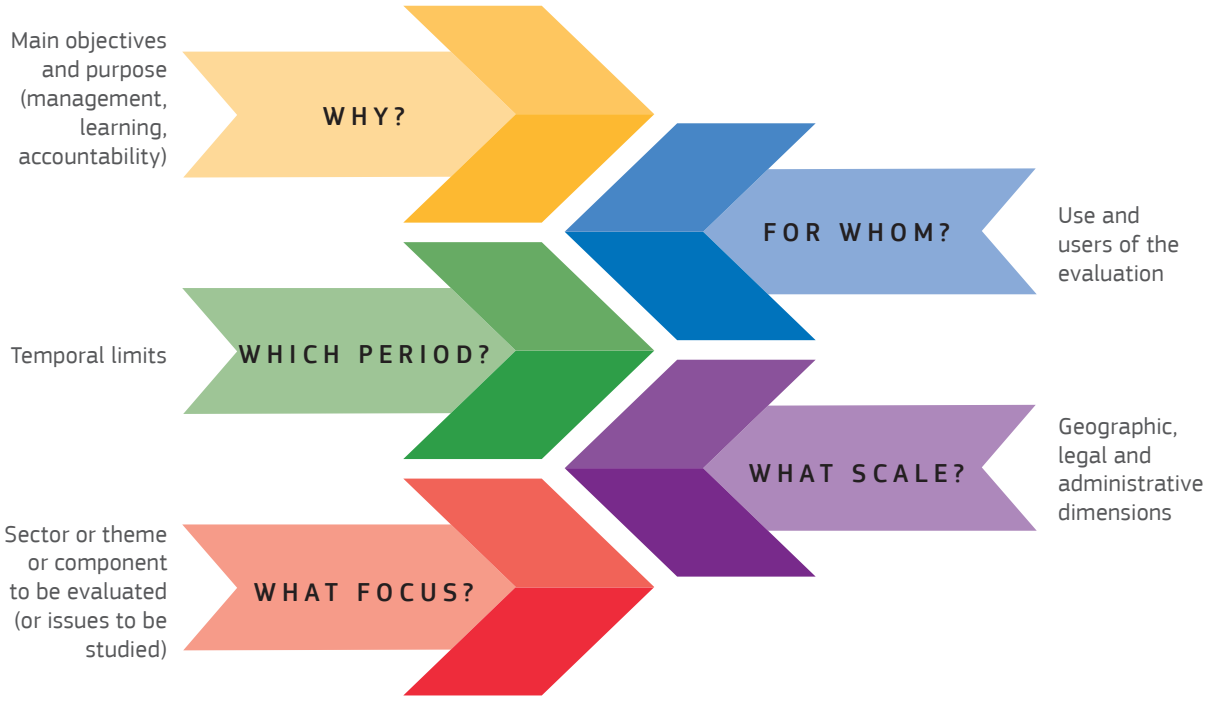
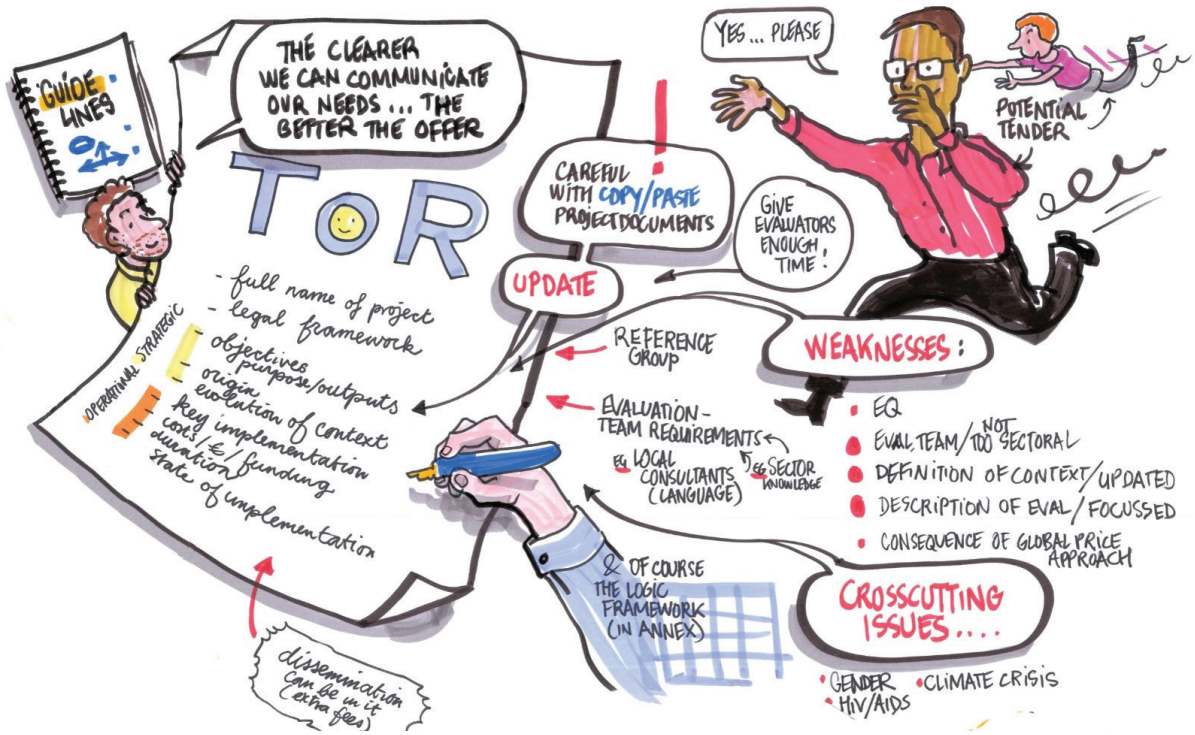


FIGURE 2.2.2 Assembling the terms of reference



2.2.3 Budgeting an evaluation

Setting the right budget for an evaluation is important. If the figure is too low, the scope and quality of deliverables are likely to be compromised; if it is too high, value for money or cost-effectiveness is compromised. Budgeting an evaluation is often challenging, as **there is no one-size-fits-all solution**; the cost of an evaluation varies significantly depending on variables such as the size and duration of the intervention, its scope and complexity, geographical coverage, size and nature of the stakeholders/target population/beneficiaries, quality of monitoring systems in place, and the proposed evaluation scope and methodology (DG NEAR, 2016).

Tools to support evaluation managers in setting appropriate evaluation budgets have been developed. These tools are based on a simple logic, which essentially breaks a given evaluation down into its constituent tasks as identified in the corresponding ToR; Table 2.2.1 shows a simpified version of this breakdown.

Other basic information that will be needed for budget estimation includes a clear understanding of the scope of the evaluation (geographical, temporal, regulatory); the number of phases (including the key tasks to be carried out by the evaluation team during each phase); a general idea of the range of stakeholders to be interviewed, including field visits; an initial idea of the planned data collection tools to be used by the team (e.g. surveys, case studies, counterfactual impact assessments, cost-benefit analyses); the number of reports and other deliverables foreseen (including any dissemination products such as videos, infographics etc.); and the number and category of experts to be mobilised, as well as the average prevailing prices per category of expert. Armed with this information – which is in any case required to draft the ToR – the evaluation manager can arrive at a relatively robust estimation of the cost of the planned evaluation.

2.2.4 Drafting the terms of reference

Submitting a complete and useful ToR can be a daunting prospect. To simplify the process, **ToR templates and specific guidance for intervention-level evaluations** are available to staff on the EU *intranet*. This section discusses guidance for drafting a ToR for evaluation of interventions under the *Services for EU's External Action 2023* framework contract; other templates are available on the *SEA 2023* Intracomm page (e.g. guidance for the drafting of ToR for budget support evaluations – intervention level.

Additionally, past **ToR (and reports) from evaluations done by partner organisations** are accessible to users through the function search (these are currently available in the *EVAL Module* legacy and will become available in EVAL OPSYS in 2025).

Box 2.2.2 presents the contents of the intervention-level evaluation ToR guidance in line with the standard template; Table 2.2.2 is a handy checklist to make sure the ToR is fully completed. The remainder of this section provides some **tips, techniques and conceptual information for preparing the ToR**. It aims to supplement, not reiterate, the detailed guidance provided in the ToR template itself and on OPSYS.

GOOD PRACTICES

- Be able to state the focus of the evaluation in a single sentence.
- Be clear and succinct; summarise and tailor, rather than reiterate, previously written generic material.
- Keep the writing factual and free of judgement, particularly with regard to the intervention and its results and performance.
- Provide accurate descriptions rather than interpretations of the intervention logic.
- Present information as set out in the ToR outline.
- Ensure consistency between the different sections of the ToR.
- Keep the background description to the minimum of what is necessary to understand the context; there is no need for a long exposé copy and pasted from a previous report.

TABLE 2.2.1 Template for budgeting an evaluation: calculating expert person days

	Expert level[1]			Comments
	Senior	Inter-mediate	Junior	
Initial desk study				Min. 3 days for task leader (TL)
Kick-off				Min. 1 day for each participant expert
Initial interviews				Consider max. 4 interviews per expert/day
Further desk study				Depends on number and size of secondary sources
Reconstruct logframe/ToC				Min. 2 days for TL, other experts need to be involved
Methodology				Min. 2 days for TL, other experts may need to be involved
Evaluation matrix				Min. 2 days for TL, other experts need to be involved
Finalise eval. questions				Min. 1 day for TL, other experts need to be involved
Develop tools				Min. 2 days for the TL, but they can be many more
Write inception report				Min. 3 days for TL, other experts need to be involved
Finalise inception report after comments				Min. 2 days for TL, other experts may need to be involved
Desk phase				If needed; depends on number and size of secondary sources
Interim report				If needed, min. 3 days for TL plus involvement of further experts
Schedule interviews				≤ 4 days, depending on number of interviewees and travel
Field missions				Consider each location separately; add as many rows as needed; include travel time
Field debrief				Min. 1 day for each participant in the field
Wrap-up				Min. 4 days for TL plus a few days for each member
Final report				Min. 7 days for TL plus substantial days from the other members
Finalise final report after comments				Min. 3 days for TL, other experts may need to be involved
Dissemination products				Depends on products selected
Dissemination seminar				≥ 1 day, depending on location, number of days and type of event (online or in person)
Total				

[1] Or as specified in the general conditions of the corresponding framework contract.

BOX 2.2.2 Outline as per current ToR guidance for intervention-level evaluations

PART A

1 Background information

1.1 Relevant country [region/sector] background

1.2 The intervention[s] to be evaluated

1.3 Stakeholders of the intervention

1.4 Previous internal and external monitoring (including ROM), evaluations and other studies undertaken

2 Objective, purpose, and expected results

2.1 Global objective of the evaluation

2.2 Specific objectives of the evaluation (including evaluation criteria and indicative evaluation questions)

2.3 The requested services including suggested methodology

2.4 Required outputs

3 Logistics and timing

4 Requirements

5 Reports/deliverables

5.1 Use of the Funding & Tenders Portal by the evaluation contractors and experts, and of EVAL OPSYS by the evaluation manager

6 Monitoring and evaluation

6.1 Content of reporting

6.2 Comments on the deliverables

7 Practical information

Annexes

I Logical framework matrix (logframe) of the evaluated interventions

II Information that will be provided to the evaluation team

III Evaluation criteria

IV Evaluation matrix

V Structure of the reports

VI Quality assessment grid

VII Planning schedule

[For evaluations of budget support programmes two additional annexes are foreseen: (i) Example of intervention logic diagram for a budget support programme; and (ii) Specificities and description of the five-level model of the intervention logic of a budget support programme]

PART B

1. Benefiting zone
2. Specific contracting authority
3. Specific contract language (and location)
4. Start date and period of implementation
5. Expertise
6. Incidental expenditure
7. Lump sums
8. Expenditure verification
9. Other details

TABLE 2.2.2 Checklist for evaluation ToR completion of intervention-level evaluations (other than budget support)

	Item	ToR reference	Y	N
	CONTEXT OF THE EVALUATION (PART A, SECTION 1. AND ANNEXES)			
1	Have you provided relevant **contextual background**?	1.1		
2	Have you provided **concise background information on the intervention(s)** and its (their) evolution during the period under evaluation (in past tense)?	1.2		
3	Have you described the **intervention logic** or **theory of change** underpinning the intervention(s) to be evaluated?	1.2.2		
4	Have you described the **key stakeholders** of the intervention(s) to be evaluated?	1.3		
5	Have you **annexed the most recent logframe(s)**?	Annex I		
6	Have you **summarised results from previous evaluations or monitoring/ROM missions** (even if financed by other agencies)?	1.4		
	EVALUATION MANDATE AND STRUCTURING (PART A, SECTION 2.)			
7	Have you defined the global and specific **objectives** of your evaluation (why the evaluation is needed, what purpose the results will be used for)?	2.1, 2.2		
8	Have you considered the **DAC evaluation criteria** based on evaluation type and objectives, e.g., by eliminating those that are not essential, and justifying this?	2.2		
9	Have you developed a set of **Indicative Evaluation Questions**, organised in a meaningful way (by DAC+EU criteria or by transversal and/or thematic clusters)?	2.2.1		
10	Is the total **number of Evaluation Questions** between 5 and 10?	2.2.1		
11	Are **most of your questions open ended** and focused on 'why' and 'how'?	2.2.1		
12	Do your Indicative Evaluation questions **refer to gender-, age- and disability-disaggregated information** (where relevant)?	2.2.1		
13	Is there clear **coherence** between the **objectives** of the evaluation (2.1 and 2.2), the **evaluation criteria** (2.2) and the **Indicative Evaluation Questions** (2.2.1)?	2.1, 2.2, 2.2.1		
14	Have you defined who will be the **key users** of your evaluation?	2.2.2		
15	Have you **considered your evaluation phases**, even in relation to the possibility of doing fieldwork?	2.3.1		
16	Have you **included an evaluation dissemination phase**?	2.3.1		
17	Did you choose **appropriate Reference Group members**? This will be between three and six participants selected from among your colleagues and other stakeholders.	2.3.3		
18	Did you ensure consistency in the **outputs and deliverables** (terminology, timing, etc.) across evaluation phases & all relevant sections (2.3.1, 2.4, 5, 6 & Annex V)?	2.3.1, 2.4, 5, 6, Annex V		
	EVALUATION TEAM (PART B)			
19	Have you described the minimum requirements for the team of experts as a whole?	Part B, 6		
20	Does your ToR assign a senior professional evaluator as **team leader**? [Recommended]	Part B, 6		
21	Have you defined the **required functions and their expected category**? [Optional]	Part B, 6		
22	Have you defined the **expected number of working days per required function**? [Optional]	Part B, 6		
	CROSS-CHECKING WITH OTHER ELEMENTS OF THE REQUEST FOR SPECIFIC CONTRACT			
	Did you ensure that your **evaluation grid** and **organisation and methodology** are aligned with the requirements of your ToR?			

SOURCE: *Guidance to the SEA ToR template for intervention level evaluations*; text has been lightly copy edited for consistency and syntax.

- If available, present the intervention logic in a graphic format.
- Be realistic in terms of the time and resources needed to adequately carry out the evaluation.

> **NOTE:** *To determine the minimum number of working days required for the evaluation team, list all the evaluation steps to be performed in the different evaluation phases (see Table 2.2.1), and then identify the minimum number of days for each category of expert for each step, including the number of days required in the field.*

- Do not underestimate the importance of evaluation expertise within the evaluation team.

> **NOTE:** *Being a thematic/sector expert with experience in evaluations is very different from being an expert in evaluation with sector/thematic experiences. As specified in the evaluation ToR guidance, a team leader should preferably be an expert in evaluation methodologies with sector experience rather than vice versa.*

> **SEE:** How-to Guide on the evaluation team profile *on the Evaluation wiki.*

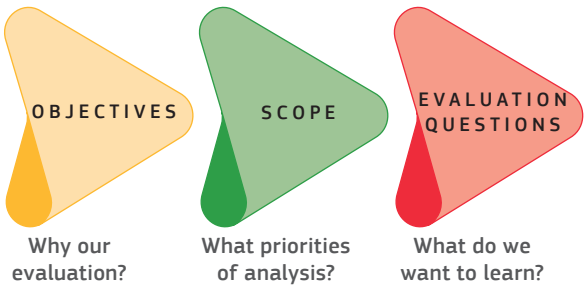
KEY CONCEPTS AND CONSIDERATIONS

Following are brief discussions of some perhaps unfamiliar technical aspects of evaluation that will be helpful in preparing the ToR. Further explication of these is provided in Chapter 3.

> **NOTE:** *Also see the discussion of evaluation timing in* Section 1.2.

Intervention logic. A valid intervention logic is fundamental to the evaluation process. It describes the expected logic of the intervention or the chain of events that should lead to the intended change. It can be presented as a narrative description and/or as a diagram summarising how an intervention is expected to deliver results. The intervention logic identifies the causal links between the outputs and the outcomes, and between the outcomes and the impact – also known as the results chain – as well as the key assumptions underpinning that change process.

FIGURE 2.2.3 Formulation of the evaluation questions



> **SEE:** Subsection 3.2.3.

Evaluation criteria. The six evaluation criteria established by the Development Assistance Committee (DAC) of the Organisation for Economic Co-operation and Development (OECD) are a de facto standard in evaluation worldwide, capturing key aspects of a strategy, policy, instrument, modality, intervention or group of interventions. Within the EU, these evaluation criteria – relevance, coherence, effectiveness, efficiency, impact and sustainability – are joined by a seventh EU-specific evaluation criterion: EU added value.

All six OECD DAC criteria may not need to be examined in every evaluation; in fact, the OECD DAC principles of evaluation emphasise that the criteria should be applied thoughtfully and not mechanically. The evaluation manager should carefully consider exactly what needs to be measured in the particular evaluation.

> **SEE:** Subsection 3.1.2.

Indicative evaluation questions. These are the foundation of the evaluation and flow directly from its objectives and scope (see Figure 2.2.3); they will tell the evaluation team what needs to be looked at and will be structured around the intervention logic and the evaluation criteria. In general, there should be between 6 and 10 evaluation questions.

> **SEE:** Subsection 3.1.3, *Tool #47 in* Better Regulation Toolbox *(EC, 2023), the discussion of* evaluation questions *on Capacity4dev's Evaluation methodological approach wiki and the* How-to Guide on evaluation questions *on the Evaluation wiki.*

These indicative questions are very much a work in progress; they are called 'indicative' because they will **not be set in stone until the end of the inception phase** (see Subsection 2.3.3).

The questions should be written as **simply and precisely** as possible. The evaluation manager should start from the basics: why is this evaluation needed, and what should be learned from it? Because every evaluation is unique, copying questions from other evaluations is not advised. The reference group members should be involved in defining the evaluation questions.

> **SUGGESTION:** *The evaluation manager may want to ask the framework contractors to further refine the indicative evaluation questions in their offers. This is a good way to check if they understand the objective of the requested evaluation. The evaluation manager could also ask for a preliminary evaluation matrix.*

Gender. Gender- and power-neutrality do not exist; actions affect women and men differently, positively or negatively, and their respective powers are key elements influencing this impact. All evaluations should therefore adopt a gender- and rights-sensitive approach, and evaluators are called on to play a key role in understanding and informing to what extent the evaluation contributes to the EU commitment to gender equality and empowering women and girls and to a human rights–based approach to development cooperation.

> **SEE:** Box 1.3, Box 3.2.1 *and discussion under* Subsection 3.1.2 *for more information on gender-responsive evaluation. Also see the EC's* Gender Equality Strategy Monitoring Portal.

Complex settings, including hard-to-reach areas and contexts affected by fragility, conflict and violence. A hard-to-reach area is one that is difficult to access because of conflict, human-engineered or natural disasters, or other physical, logistical, security or health-related obstacles (EC, 2019). Given their fluidity and unpredictability, non-traditional evaluation approaches are often used in such settings – including directly engaging the implementing partners and other stakeholders in the evaluation.

> **SEE:** Box 2.3.2 *for more information.*

2.2.5 Managing the contractual procedures

SUBMIT THE ToR

Once the ToR is ready and the necessary funding has been secured, the evaluation manager launches the request for services by registering it with OPSYS.

> **OPSYS:** *The evaluation manager launches the request for services by (i) accessing the EVAL wiki on Creation of an Evaluation; and (ii) accessing the OPSYS wiki on the Request for Services – Initiation, Preparation, Verification and Authorisation.*

RECEIVE AND ASSESS PROPOSALS

Once the request for services is launched, the framework contractors confirm by the next working day their intent to submit an offer.

BOX 2.2.3 Checklist for assessing the quality of a proposal

Capacity

- Expertise in evaluation methodology and working experience in evaluation
- Demonstrated ability to carry out the evaluation
- Technical and sectoral knowledge and expertise
- Capacity to address essential cross-cutting thematic issues (e.g. gender equality, environment)
- Experience in development cooperation, and EC cooperation in particular
- Experience in the partner region, similar countries and/or the partner countries
- Adequate language capacity

Understanding

- Understanding of the ToR
- Understanding of the context

Management

- The team leader should be an expert in evaluation methodology and should have completed at least one evaluation as a team leader previously
- Proposed individuals can successfully complete their tasks as planned in the time schedule
- Clear sharing of responsibilities and adequate leadership skills for effective team management

NOTE: *The default term for proposal submission is 15 days as specified in the ToR, but the evaluation manager can extend this given an evaluation's complexity or the difficulty of locating a particular expertise. Only in cases of extreme necessity should the deadline be shortened.*

The evaluation manager receives the technical and financial proposals prepared by the potential contractors and checks that they include:

- an understanding of the ToR;
- an indicative methodological design;
- a planned schedule;
- a description of the team members' responsibilities, curricula vitae (CVs) and signed statements of absence of conflict of interest.

NOTE: Independence *and objectivity are essential to the credibility of evaluation findings, conclusions and recommendations. Evaluation team members therefore need to be independent from any organisations that have taken part in the design and implementation of the intervention to be evaluated.*

The evaluation manager assesses the quality of the proposal(s) (see Box 2.2.3) and checks that the proposed human and financial resources are adequate and in line with the requirements laid down in the ToR. The assessment is carried out based on the scoring system documented in the ToR.

ENGAGING WITH THE CONTRACTOR

Once the best offer has been selected and officially announced, the evaluation manager engages with the contractor and the evaluation team to share additional documentation and information and plan the kick-off meeting.

SECTION 2.3

Inception phase

During the inception phase, the evaluation manager, the reference group and the contractor/evaluation team arrive at a **clear, shared understanding of what is required by the evaluation** – including its structure and methodology – using the terms of reference (ToR) and the winning contractor's proposal as their starting point.

The inception phase starts as soon as the evaluation team is engaged. It can continue for a few weeks for simple evaluations and up to a month or more for complex or strategic evaluations. Its output is a **final inception note/report** that delineates:

- the reconstructed intervention logic;
- the key stakeholders;
- the final evaluation questions, including their judgement criteria and indicators;
- the evaluation methodology, including risks and mitigation measures;
- how evaluation findings will be disseminated;
- ethical considerations;
- the work plan.

Several activities are performed during the inception phase to arrive at this understanding and be able to document it. These include the following, which are not necessarily in chronological order:

- review of background documents;
- kick-off meeting for the evaluation team with the relevant European Union (EU) service, the evaluation manager and/or reference group members;
- subsequent meetings with the evaluation team, evaluation manager and the reference group;
- interviews with key stakeholders;
- research and analysis.

The remainder of this section is structured in terms of outputs needed for the inception note/report rather than tasks, since these will vary. Hyperlinks are provided to more substantive, detailed discussions in Chapter 3 as relevant.

2.3.1 Gathering data and defining the scope

The evaluation manager provided an outline of the evaluation's scope (geographic coverage, period under consideration, regulatory framework or basis etc.) in the ToR (see Subsection 2.2.2). In the inception phase, those aspects are revisited and refined – and, importantly, any limitations identified – through research and discussion conducted by the evaluation team in coordination with the evaluation manager and the reference group.

> **NOTE:** *Clearly identifying the scope ensures clarity about the evaluation mandate and expectations, allowing the evaluation team to focus precisely on priorities and avoid wasting resources.*

If deemed relevant by the evaluation manager, the evaluation scope can be extended to include related EU policies and interventions, the partner country's related policies and interventions, the partner country's poverty reduction strategy or other sector policies or interventions, or other donor interventions as relevant.

DOCUMENTATION

The evaluation manager, assisted by the reference group, collects **relevant documentation pertaining to the** evaluand **and shedding light on its context**, as listed in Box 2.3.1, and provides them to the evaluation team.

> **NOTE:** *The evaluation team will supplement these documents with others identified through independent research and during interviews with relevant informed parties and stakeholders.*

The evaluation team consults all relevant management and monitoring documents/databases to acquire a comprehensive knowledge of the evaluand, especially with regard to its **resources** (planned, committed and disbursed), **progress** of outputs/outcomes, and contact information for **potential informants**.

> **NOTE:** *This can be a time-consuming process, particularly as the online platforms and tools may work slowly with low-bandwidth networks.*

BOX 2.3.1 Information to be provided to the evaluation team

The following is an indicative list of the documents the evaluation manager makes available to the evaluation team shortly after contract signature; not all of these will be available in all cases.

- Legal texts and political commitments pertaining to the intervention(s) to be evaluated
- Country strategy paper and country/regional/thematic multi-annual indicative programming (MIPs and RIPs) documents (and equivalent) for the period covered
- Documents setting out the policy framework in which the intervention takes place (EU development and external relations policy, EU foreign policy, country strategy paper)
- Partner government strategy (e.g. poverty reduction strategy paper)
- Relevant national/sector policies and plans from national and local partners and other donors
- Ex ante evaluation, if applicable
- Intervention identification studies
- Intervention feasibility/formulation studies
- Intervention financing agreement and addenda
- Intervention quarterly and/or annual progress reports, and technical reports
- European Commission (EC) results-oriented monitoring (ROM) reports, and other external and internal monitoring reports covering the intervention
- The intervention's midterm evaluation report and other relevant evaluations, audit and reports
- Relevant documentation from national/local partners and other donors
- Calendar and minutes of all meetings of the intervention steering committee
- EC guidance on Evaluation with Gender as a Cross-cutting Dimension, evaluation questions, evaluation tools and evaluation dissemination
- Any other relevant documentation

KICK-OFF MEETING

The evaluation manager and the evaluation team should have a kick-off meeting (remote or face-to-face) **as early as possible** in the inception phase. If possible, members of the relevant EU service and/or reference group should also attend. The objectives of this meeting are:

- to arrive at a clear and shared understanding of the scope of the evaluation, its limitations and feasibility – this includes ensuring that any deviations from the ToR are later highlighted and documented in the inception and final reports;
- to discuss the evaluation methodology;
- to ensure that compulsory templates are used by the evaluation team;
- to highlight the importance of the quality of the evaluation matrix for the inception and subsequent phases;
- to ensure that sufficient time is allowed to validate the field methodology outlined in the inception report before undertaking field activities;
- to convey any other relevant points such as communication channels to be established between the parties, how quality will be assured and how to address any problems that may arise.

The kick-off meeting is an opportunity for the participants to discuss the methodology the evaluation team outlined in its proposal and to flag any concerns.

INITIAL INTERVIEWS WITH KEY STAKEHOLDERS

The evaluation manager, the EU service and/or the reference group members **facilitate the evaluation team** in setting up initial interviews with the main stakeholders (e.g. by providing a formal letter of introduction for the team to use when contacting potential interviewees).

> **NOTE:** *If a stakeholder map was annexed to the ToR, the evaluators will refine and finalise this during the inception phase in order to identify the key informants to be interviewed/surveyed. This updated map will be part of the inception note/report.*

The stakeholders to be interviewed during this phase should have a sound knowledge of the evaluand and be in a position to provide:

- a holistic perspective on the evaluand in terms of its rationale, evolution, progress to date and any major obstacles encountered;
- information to complete the reconstruction of the intervention logic and frame the evaluation questions;
- facilitate and identify other useful sources for data collection.

2.3.2 Tackling the intervention logic

A critical task for the evaluation team during the inception phase is to reconstruct the evaluand's intervention logic – meaning to **validate its original design or modify it to reflect any conditions that have changed in the interim**. This demanding and exacting work supports finalisation of the evaluation questions and design of the evaluation methodology.

> **SEE:** Subsection 3.2.3; *discussion of the* Intervention Strategy *on Capacity4dev's Evaluation methodological approach wiki; and Tool #46 in the* Better Regulation Toolbox *(EC, 2023) for more on what is entailed in reconstructing the intervention logic.*

Why? In many cases, the quality of the original intervention logic may be weak, or the context may have evolved to such an extent that the original is now obsolete. In other cases, the intervention logic has not been updated to reflect changes made during implementation. In these cases, the evaluation team will need to reconstruct the intervention logic to ensure it **adequately captures the planned change process** – the hierarchy of expected results (outputs, outcomes, impact etc.); and induced outputs in the case of budget support evaluations) – and the assumptions deemed necessary for the intervention to deliver as planned.

In the absence of any intervention logic (unlikely but possible), the evaluation team will need to draw on available documentation and initial interviews to reconstruct it from scratch.

How? Analysis of the intervention logic covers:

- the intervention context at launch, including the existing opportunities and constraints at that time;
- the needs to be met, problems to be solved and challenges to be addressed;
- assessment of the rationale for the proposed approach to addressing the identified needs, problems or challenges;
- the logical hierarchy between the different results levels (output-outcome-impact) and assumptions – for example, contextual, operational, hypothetical, environmental – underpinning those results;
- the nature of inputs and activities.

Once the analysis has been performed on the basis of available documents, the evaluation team will consult with key informants to corroborate the reconstructed intervention logic and fill any gaps.

Output. The evaluation team should prepare a graphic representation of the reconstructed/finalised intervention logic.

> **NOTE:** *Typically, this is a logical framework matrix (logframe), but other methods of presentation are possible and sometimes preferable, as discussed in Subsection 3.2.3 and Subsection 3.2.4.*

2.3.3 Refining the evaluation questions

The evaluation questions are the foundation upon which the evaluation – its methods, modes and means – is built (Figure 2.3.1). During the inception phase, the evaluation team, in collaboration with the evaluation manager and (ideally) the reference group, refines and finalises the evaluation questions based on:

- the indicative questions contained in the ToR;
- the reconstructed intervention logic;
- their reasoned coverage of the evaluation criteria.

> **SEE:** *Subsection 3.1.3 for guidance. There should be no more than 10 evaluation questions in all, and up to 12 in the case of budget support.*

Each indicative evaluation question is assessed, taking account of:

- the origin of the question and the potential utility of the answer;
- the clarity of its formulation;
- any foreseeable difficulties and feasibility problems in answering the question.

FIGURE 2.3.1 **Focusing an evaluation through evaluation questions**

For each evaluation question, the evaluation team identifies corresponding judgement criteria, indicators, targets, sources of information and data collection tools, and methods of data analysis (Figure 2.3.2). This work might require additional meetings with relevant delegations/units and, in country, at the delegation, with the national coordinator and with the partner country authorities and representatives of civil society. The scope of this activity depends on the scope of the evaluation.

> **SEE:** Subsection 3.1.3 *for information about judgement criteria and indicators.*

The evaluation questions must be agreed with the reference group. Ideally, this should happen before finalising the ToR but, if not possible, during the inception phase.

2.3.4 Finalising the evaluation methodology

Once the evaluation questions, judgement criteria and indicators have been finalised, the evaluation team should adapt, refine and finalise the evaluation methodology set out in its initial proposal. The methodology should do the following.

- Clearly explain the **conceptualisation** of the evaluation and the **approach** to be used to try to understand the extent of the change and the reasons why it happened.

  > **SEE:** Section 3.2 *for more on evaluation approaches.*

- Specify the **tools and methods** that will be used to collect the evidence to respond to the evaluation questions.

  > **SEE:** Subsection 3.3.2.

- Explain **how individuals, groups and organisations will be consulted with**, especially for complex evaluations, including those to be conducted in hard-to-reach areas and/or contexts affected by fragility, conflict and violence (see Box 2.3.2).

  > **SEE:** Subsection 3.3.5 *and* Subsection 4.7*; DG INTPA (2021); and* EvalCrisis Home *on the Capacity4dev website.*

- Identify the **limitations and risks** to be faced during the evaluation and define corresponding mitigation measures.

  > **NOTE:** *This analysis should identify any potential negative impacts, the likelihood of their occurring and how they might be avoided and/or mitigated.*

- Be **gender sensitive**, including consideration of the use of sex (and age)–disaggregated data and assessing if and how the evaluand has contributed to progress on gender equality.

  > **SEE:** Box 1.3, Box 3.2.1 *and discussion under* Subsection 3.1.2 *for more information on gender-responsive evaluation.*

- Spell out the **ethical rules** that will underlie the evaluation, both general – avoiding harm, addressing conflicts of interest, informed consent, confidentiality etc. – and pertaining to local governance and regulations.

  > **SEE:** Chapter 4.

FIGURE 2.3.2 Evaluation question aspects and considerations

Rationale	Why was the question asked?
Scope	What does the question cover?
Judgement criteria	How will the merits or successes be assessed?
Indicators	What data will help assess the merits or successes?
Targets	What level or threshold is to be qualified as a success?
Analysis strategy	What type(s) of analysis are to be undertaken?
Tools and sources of information	Where will the data come from, and how will the data be collected?

BOX 2.3.2 Evaluation and hard-to-reach areas and contexts affected by fragility, conflict and violence

Logistics and security risks make travel to some areas of the world challenging – and sometimes impossible. Collectively, these regions are defined as hard-to-reach areas and include:

- locations where natural or human-made disasters recently occurred;
- places where geographic, logistic or health-related considerations make access difficult;
- contexts affected by fragility, conflict and violence.

This last category is particularly significant. Around 2 billion people and half of the world's poor live in fragile or conflict-affected states (Corral et al., 2020; Hoogeveen and Pape, 2020). Upwards of half of all global development cooperation funding – including from the EU – is committed in such regions.

A traditional approach to evaluation in hard-to-reach areas is destined to fail, as few professional evaluators are available to travel to these areas, and security risks make conventional field missions unrealistic.

Ways around these difficulties do exist and should be explored during evaluation planning. The following guidance is oriented specifically towards contexts of fragility, conflict and violence, but much is applicable to the broader category of hard-to-reach areas in general.

While traditional evaluation focuses primarily on assessing the intended and actual results of an intervention, **conflict-sensitive evaluation** will require that particular attention be placed on (i) an understanding of the context as it changes over time and (ii) measuring the interaction between the intervention and the context.

Some key questions to think about when designing an evaluation methodology in contexts of fragility, conflict and violence include:

- Do the evaluation design and evaluation matrix have conflict-related indicators to track the evolution of the context and how the intervention is affecting or has affected the conflict? (See Goldwyn and Chigas, 2013.)
- Do the methodology and tools recognise which inter-group conflict stands out as destructive? What are the factors that divide (cause tension between) those groups? What factors connect them (or help them manage conflict)?
- Does the evaluation matrix contain questions on what disputes arose during the intervention? What were their underlying causes? Were the disputes addressed and, if yes, how? Were any disputes avoided and, if yes, how? What roles did different actors play in dispute resolution, if any?

Look explicitly at the **gender and conflict links** through conflict analysis and gender analysis of conflict. These analyses are aimed at helping understand gender dynamics in the context of fragility. This can include:

- What are the different needs and aspirations of women and men and boys and girls in the conflict situation?
- How are the respective gender roles of women and men and boys and girls affected by the conflict?
- What roles are they playing in bringing about a peaceful resolution to the conflict?

For additional guidance and information, see EC (2019); Conflict Sensitivity Consortium (2012); Hassnain, Kelly and Somma (2021); and Woodrow, Oatley and Garred (2017).

- Operationalise intercultural elements in the evaluation questions.

NOTE: *See* Box 4.2.

For a **strategic evaluation or an evaluation covering multiple interventions**, the methodology also needs to include a proposed representative sample of interventions to be analysed in greater detail to inform the assessment of performance and results/sustainability. The selection of this sample should be underpinned by a clear methodology, including the criteria to be used. For **budget support evaluations**, the specific methodology developed for that purpose should be used.

SEE: *The Annex to this handbook for more on budget support evaluations.*

2.3.5 Drafting the inception note/report

Based on analysis and exchanges with the evaluation manager, the reference group and other key stakeholders, the evaluation team drafts the inception note/report.

> **NOTE:** *The content of the inception note/report is outlined in an annex to the original ToR. See Introduction to the Terms of Reference and Guidance Notes for Evaluations of Interventions for guidance.*

The structure and format of this note/report is at the team's discretion; the document should be no more than 20 pages long (excluding annexes) and must include the following:

- **introduction**, including a short description of the evaluation's context, objectives, focus and intercultural elements;
- **reconstructed intervention logic**, presented in a logframe or alternative format;
- **stakeholder map**, indicating the key stakeholders and their relations with the evaluand;
- **evaluation methodology**, including:
  - an overview of the entire evaluation process
  - tools and methods for data collection
  - consultation strategy, including sampling approach, if relevant
  - proposed case studies and field missions;
- **analysis of risks** related to the evaluation methodology and mitigation measures;
- **ethics rules**;
- **finalised evaluation questions**, presented using the evaluation matrix template (Part A);

> **SEE:** Evaluation matrix *discussion below for guidance on preparing this matrix.*

- **work plan**, typically presented as a Gantt chart with accompanying text;

> **SEE:** Work plan *discussion below for more information.*

- **annexes**, including the ToR, a list of documentation consulted and a list of people met with/interviewed during the inception phase.

An outline of the data collection tools should also be included in the inception note/report, but it may not be feasible in all cases. The dissemination strategy should also be described, or at least outlined, in the inception report.

> **SEE:** Section 2.6 *for information on dissemination strategies.*

EVALUATION MATRIX

The evaluation questions are summarised in Part A of the evaluation matrix (Figure 2.3.3 and Figure 2.3.4), which is then included as part of the inception report/note.

> **NOTE:** *This matrix is used throughout the evaluation and is updated and submitted by the evaluation team along with every report provided during the evaluation (i.e. the inception note/report, the desk report and the final report).*

The first lines of Part A capture the text of the question, plus a brief explication of why the question was asked and a clarification of terms used if necessary. The subsequent lines develop the chain of reasoning by which the evaluation team plans to answer the question, beginning with each judgement criterion and its associated indicators. The chain is described through a series of steps to be taken by the evaluation team in order to:

- inform on change in relation to the selected indicators;
- assess causes and effects;
- assist in the formulation of value judgements.

These steps need to be associated with the corresponding information sources and data collection tools.

Part B, the evidence log, is completed during the synthesis phase. It is a key tool for the evaluation team, as it allows them to track the data they are capturing for each indicator and to assess the quality of that data/evidence.

FIGURE 2.3.3 Evaluation matrix

PART A – EVALUATION DESIGN

Evaluation Question 1:				
EVALUATION CRITERIA COVERED				
JUDGEMENT CRITERIA (JC)	INDICATORS (IND)	INFORMATION SOURCES		METHODS / TOOLS
		PRIMARY	SECONDARY	
JC 1.1 –	I 1.1.1 –			
	I 1.1.2 -			
	I 1.1.3 -			
JC 1.2 –	I 1.2.1 -			
	I 1.2.2 -			
	I 1.2.3 -			

To be completed during inception phase

PART B – EVIDENCE LOG

Ind	Baseline data	Evidence gathered/analysed	Quality of evidence
I 1.1.1			
I 1.1.2			
I 1.1.3			
I 1.2.1			
I 1.2.2			

To be completed during synthesis phase

One set of tables (Parts A and B) should be used for each evaluation question (see Figure 2.3.4), adding or deleting as many rows as needed to reflect the selected judgement criteria and indicators.

WORK PLAN

The work plan clearly describes the activities to be carried out in the following phases of the evaluation, including field missions, surveys, focus group discussions, case studies, interviews etc. It should also indicate the human and technical resources needed to conduct these activities. In establishing the work plan, the evaluation team should be sure to include sufficient buffer time to allow for delays, addressing of quality issues etc.

When fully drafted, the inception note/report is presented to the reference group in either a remote or face-to-face meeting.

2.3.6 Approving the inception note/report

The evaluation manager is responsible for coordinating review of and feedback on the draft inception note/report. To this end, the evaluation manager:

- **distributes the report** to the reference group and follows up to ensure that the members of the group provide their feedback on time;
- **consolidates all comments** and submits them to the evaluation team, setting an agreed-upon deadline for delivery of the revised note/report.

Table 2.3.1 and Table 2.3.2 present checklists for the evaluation manager to use in determining the completeness/quality of the inception note/report for intervention- and strategic-level evaluations, respectively.

FIGURE 2.3.4 Sample partially completed evaluation matrix

Evaluation Question 1: To what extent does the support provided by the evaluand correspond to the needs, priorities and capacities of the partner countries and their supported authorities?				
EVALUATION CRITERIA COVERED	Relevance			
JUDGEMENT CRITERIA (JC)	INDICATORS (IND)	INFORMATION SOURCES		METHODS / TOOLS
		PRIMARY	SECONDARY	
JC 1.1 – The intervention activities responded to the needs and priorities of the partner countries and their supported authorities	I 1.1.1 – The extent to which partner countries' needs and priorities are incorporated into intervention design	Notes from interview with implementing partners on incorporation of priorities and needs of partner countries during design phase of the intervention	Inception report Initial country reports Progress reports	Descriptive evaluand document analysis Semi-structured interviews
	I 1.1.2 -			
	I 1.1.3 -			
JC 1.2 – The activities took into account the capacities of the partner countries	I 1.2.1 - The extent to which partner countries' capacities assessment is incorporated into intervention design	Notes from the interview with implementing partners on incorporation of capacities assessment of beneficiary countries during design phase	Inception report Initial country reports Progress reports	Descriptive evaluand document analysis Semi-structured interviews
	I 1.2.2 -			
	I 1.2.3 -			

The evaluation team finalises the inception note/report based on the comments and feedback.

> **NOTE:** *It is critical that a good-quality inception report be developed prior to the launch of the next phase of the evaluation. To this end, sufficient time **must** be factored into the evaluation timeline. All too frequently, evaluation managers feel obliged to accept substandard inception reports due to time pressures. This practice is a major contributor to poor-quality evaluation reports.*

> **OPSYS:** *The evaluation team uploads the final inception note/report.*

Once the report is considered to meet the quality requirements, the evaluation manager officially approves the report and, through ARES (the Advanced Record System), issues a **formal letter** authorising continuation of the evaluation team's work.

> **OPSYS:** *The evaluation manager approves the inception note/report.*

> **NOTE:** *The work plan and evaluation questions become contractually binding upon approval of the inception note/report. Any significant deviations from these that could compromise the quality of the evaluation or jeopardise its completion within the contractually agreed-upon time frame should be immediately discussed with the evaluation manager, and necessary corrective measures undertaken.*

TABLE 2.3.1 Intervention-level inception note/report quality review checklist

	Item	Y	N
	1. CLARITY OF THE REPORT		
1.1	Easily readable and understandable (free of jargon; written in plain English or French; logical use of chapters; appropriate use of tables, graphs and diagrams)		
1.2	Appropriate length (20–30 pages, excluding annexes)		
1.3	Annexes contain (at a minimum) the original terms of reference (ToR), the evaluation matrix, a bibliography and a list of consultees; and comply with what was required and specified in the ToR		
	2. INTRODUCTION		
2.1	Provides an appropriate description of the evaluation's context, and its objectives and focus		
2.2	Explains the timing of the evaluation and its expected outputs and use		
2.3	Any departures from the original ToR are adequately explained and justified		
	3. FINALISED EVALUATION QUESTIONS WITH JUDGEMENT CRITERIA AND INDICATORS		
3.1	The total number of Evaluation Questions are equal to or lower than 10		
3.2	Questions are specific, open-ended and focused on 'why' and 'how'		
3.3	Questions are sensitive to the context, including gender, age, and disabilities issues (as relevant)		
3.4	The judgement criteria and indicators are well defined, relevant to the EQs and coherent to the objectives of the evaluation (ToR 2.1), the selected OECD/DAC criteria (ToR 2.1) and the indicative Evaluation Questions (ToR 2.2)		
3.5	The evaluation matrix is clearly articulated indicating the evaluation criteria, data sources and methods for gathering and analysis.		
	4. METHODOLOGY OF THE EVALUATION		
4.1	The IR provides a detailed overview of the evaluation process, its design and the choice of evaluation criteria		
4.2	The proposed design, tools and methods are appropriate for addressing the evaluation mandate and their relative strengths are explained		
4.3	The consultation strategy is clear and appropriate		
4.4	The structuring and organisation of the different phases of the evaluation, including planning of the missions is clear		
4.5	Methodological limitations are acknowledged, their impact on evaluation design is discussed and appropriate mitigation measures envisaged.		
	5. RISKS AND ETHICS		
5.1	The IR explains how the evaluation avoids harm; attains informed consent; ensures confidentiality and demonstrates contextual sensitivity		
5.2	The IR contains a section describing actual or potential conflict of interest affecting the evaluation team and an appropriate mitigation strategy is explained.		
	6. WORK PLAN		
6.1	A sufficiently detailed free text description of the work plan is provided in the IR		
6.2	The work plan is provided in Gantt format		

TABLE 2.3.2 Strategic-level inception note/report quality review checklist

Item	Rating	Comments, feedback and recommendations to the authors for improvement
1. STRUCTURE AND CLARITY		
1.1 The inception report is consistent with the content and requirements requested by the ToR		
1.2 The product is accessible to the relevant services and the main users as specified in the ToR (free of jargon; written in plain language; appropriate use of tables, graphs and diagrams)		
1.3 The annexes contain what is requested by the ToR – e.g. at a minimum, the original ToR, the evaluation framework, a bibliography and/or list of documents consulted, a stakeholder map and a list of consultees.		
2. PURPOSE, SCOPE AND OBJECTIVES		
2.1 The purpose and objectives of the evaluation are clearly articulated; accountability and learning aspects have been considered and it is clear to the reader why the evaluation is being undertaken		
2.2 The report explains the target audience(s) for the evaluation findings		
2.3 The scope is clear: the report explains what aspects of the intervention are included in and excluded from the evaluation; the boundaries of the scope are well justified, as are any overlaps with related policies that will be included		
2.4 Key stakeholders for the evaluation data collection been identified and, where relevant, their participation in the inception phase is explained		
2.5 Any departures from the original ToR have been adequately explained and justified		
3. CONTEXT		
3.1 The report provides a brief analysis of the geographical, sector and policy contexts which are appropriate and relevant to the evaluation scope and objectives		
3.2 The product identifies key linkages between the intervention and other relevant policies/programmes/donors; if no linkages are identified, the report justifies why other policies/programmes/donors will not be relevant to the evaluation		
3.3 The intervention logic and/or theory of change that will be used for the evaluation is clearly presented; if this was reconstructed during the inception phase, the report describes the development or consultation process		
4. EVALUATION DESIGN AND FRAMEWORK		
4.1 The evaluation methodology is clearly articulated and justified, and complies with required standards; appropriate and relevant criteria are adequately reflected in the evaluation framework		
4.2 Evaluation questions and judgement criteria have been identified and comply with required standards; they are sufficiently specific and clearly related to the evaluation purpose, scope and objectives		

(continued)

TABLE 2.3.2 Strategic-level inception note/report quality review checklist (continued)

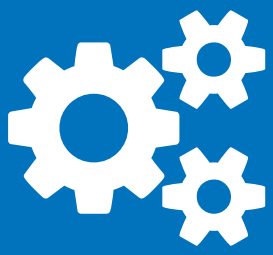
Item	Rating	Comments, feedback and recommendations to the authors for improvement
4.3 The evaluation questions and their related judgement criteria strongly derive from the results and impacts in the intervention logic or theory of change, and the related sectors, themes and instruments; the product provides a relevant and sufficient description of whether and how contextual factors (local, national and/or international) have influenced the evaluation design and planned process		
4.4 The indicators are relevant to the judgement criteria and evaluation questions; they are RACER/SMART to the extent possible and realistic – evidence is very likely to be obtained for the indicators given the context and the methods proposed		
4.5 The report justifies any changes in the evaluation questions proposed in the ToR, if relevant		
4.6 The evaluation framework will address the cross-cutting issues identified in the ToR		
5. METHODS AND DATA		
5.1 Data gathered to date as part of the inception phase (including accessing databases) are clearly presented and relevant to the evaluation objectives		
5.2 The report describes well all of the data collection methods to be applied throughout each phase		
5.3 The data analysis strategy is explained and justified; the design provides for multiple lines of inquiry and/or triangulation of data and, if not, there is a clear rationale for doing otherwise		
5.4 The methods and sequencing are appropriate for addressing the objectives of the evaluation (i.e. all of the evaluation questions)		
5.5 Primary and secondary data sources are appropriate, adequate and reliable; the report indicates the quality of data sources or whether the quality is not yet known		
5.6 The sampling strategy is described – the approach is appropriate and the sample size is adequate; this applies to all types of data sources: stakeholders, field site selection, documents		
5.7 There are adequate plans to consult with different stakeholders at all levels		
5.8 The methods and data sources will enable the collection and analysis of disaggregated data to show differences between groups		
5.9 Pre-testing the tools has been built into the data collection phase		
5.10 Both methodological and contextual limitations are acknowledged and their impact on the evaluation design discussed; limitations are acceptable and/or are adequately mitigated against		
6. ETHICAL CONSIDERATIONS		
6.1 The evaluation design includes explicit consideration of INTPA's and FPI's commitment to integrating rights-based approaches to development cooperation, and issues of equity and gender have been considered particularly in relation to the inclusion of stakeholders and participants		

(continued)

TABLE 2.3.2 Strategic-level inception note/report quality review checklist (continued)

Item	Rating	Comments, feedback and recommendations to the authors for improvement
6.2 If the evaluation participants include community members who are disadvantaged or the evaluation content includes sensitive issues (e.g. personal health issues, HIV, violence, gender inequality), the report explains how formal ethical approval at the appropriate national/ organisational level will be/has been obtained for this work and how informed consent will be managed		
6.3 The inception report explains how stakeholders who will be affected by the intervention will be provided with appropriate access to evaluation-related information in forms that respect confidentiality (beneficiary feedback)		
6.4 A clear and comprehensive plan is included to manage data responsibly; this means that data storage and protection approaches are explained, and the evaluation process demonstrates how it will uphold the privacy and confidentiality of evaluation participants/stakeholders		
6.5 Any actual or potential conflict of interest affecting the evaluation team is disclosed and an appropriate mitigation strategy is explained		
7. PLANNING, MANAGEMENT AND GOVERNANCE		
7.1 It is clear who will be undertaking the evaluation; the roles and responsibilities of the evaluation team are clearly defined; accountabilities, responsibilities and lines of communication within the evaluation team, and between the evaluation team and commissioners, are absolutely clear		
7.2 The evaluation team composition is explained in terms of both sectoral and methodological expertise; the team leader has human resource management skills and a proven track record of timely high-quality evaluations		
7.3 The team structure for the evaluation includes diverse perspectives, and the report confirms or demonstrates such perspectives will be free of control from organisational influence and political pressure		
7.4 The team's approach and plan for managing quality assurance is included and appropriate		
7.5 A sufficiently detailed work plan (including timeline and team inputs) is provided; it is feasible and includes appropriate timing for quality assurance of evaluation products (i.e. baseline report if applicable and evaluation report)		
7.6 Any risks and challenges identified within the original ToR or through the inception process have been adequately addressed		
7.7 Coordination with the relevant policies and evaluations of other donors has been considered in evaluation design (to an extent proportionate to the evaluation effort and purpose) in order to minimise burdens and transaction costs on the partner country		

NOTE: This figure is drawn from a standard INTPA form, modified to highlight the criteria on which strategic-level evaluations are to be assessed. The ratings to be used are: ***Excellent*** – the criterion was fully met (or exceeded) and there were no or few shortcomings; evaluation commissioners may use the inception report with a high degree of confidence that the design will meet the needs of the evaluation. ***Good*** – the criterion was met with only minor shortcomings; evaluation commissioners may use the inception report with confidence that the design will meet the needs of the evaluation when some improvements have been addressed. ***Needs improvement*** – the criterion was partially met with some shortcomings; decision makers may continue to proceed with commissioning the evaluation, but substantive improvements are advised to ensure that the design will meet the needs of the evaluation. ***Unsatisfactory*** – There were major shortcomings in meeting INTPA standards for inception reports; evaluation commissioners may not rely on this inception report to meet the needs of the evaluation. ***N/A*** – not applicable; the question/criterion is omitted from scoring/rating.



SECTION 2.4

Interim phase

The interim phase comprises desk and field activities during which the evaluation team (i) collects and analyses data and information to arrive at preliminary hypotheses that are then (ii) validated and/or revised on site or remotely, if travel is not possible.

During the **desk activities**, the evaluation team carries out data collection and analysis, including document review, key stakeholder interviews and other forms of data collection (e.g. surveys) and identifies any information gaps. Partial answers to the evaluation questions are formulated on the basis of existing information in line with the approved evaluation matrix. The analysis should identify the hypotheses to be tested in the field, as well as develop a methodology for field visits. The **field visits** allow the team to complete their data collection and either confirm or adapt the initial hypothesis through a process of triangulation.

The main emphasis of the phase is **data**: how to collect it and how to analyse it. A variety of data collection tools and techniques are identified in Figure 2.4.1; more detailed information can be found in Section 3.3.

Desk and field activities can be undertaken sequentially, in parallel or jointly. This section covers the two sets of activities separately, with a final subsection briefly describing how they can be combined.

FIGURE 2.4.1 **Examples of data collection tools**

2.4.1 Desk activities

The desk activities seek to:

- collect as much relevant information as possible;
- analyse the data;
- draft preliminary answers to the evaluation questions;
- identify the hypotheses to be tested in field activities.

To achieve these objectives, the evaluation team performs the following tasks.

TOOL SELECTION AND DEVELOPMENT

The tools to be used to collect data during the interim phase were identified and outlined in the inception note/report, but may need to be adapted. Data collection tools range from simple, standard instruments such as interviews, surveys and case studies, to more technical ones such as modelling or cost-benefit analysis.

NOTE: Box 2.4.1 *summarises tools frequently used by evaluators; more complete descriptions for these and others can be found in* Subsection 3.3.2.

BOX 2.4.1 Evaluation tools

TRADITIONAL

- Interviews
- Surveys
- Focus group discussions
- Case studies
- Document review
- Observation

INNOVATIVE

- Big data
- Online smartphone-based surveys
- Social media and crowdsourcing
- Geospatial technology
- Geo-Enabling Initiative for Monitoring and Supervision (GEMS)
- Story gathering/inquiry tools
- Participatory videos

For more information, see Subsection 3.3.2. For a discussion on techniques for and mitigating the risks of remote data collection (including in pandemic settings and settings of fragility and conflict), see Hassnain (2020) and Hassnain and Lorenzoni (2020b).

The choice of tools will depend on the objectives of the particular evaluation and its context (see Box 2.4.2), but should include an appropriate mix of tools to:

- collect both quantitative and qualitative data;
- allow for cross-checking (triangulation) of information from different sources;
- align with the planned time frame and available resources.

BOX 2.4.2 Key criteria for selecting a mix of evaluation tools

- Specific functions and ability to be implemented
- Need for specific data (check availability and reliability in advance)
- Necessary resources for using the tools
- Necessary time for preparing and using the tools
- Availability of qualified and suitably skilled experts (good knowledge of national languages and cultures, field experience, experience with specific tools)

Each tool is developed through a preparatory stage which covers all or part of the following items:

- list of questions and steps of reasoning to be addressed with the tool;
- technical specifications for implementing the tool;
- foreseeable risks that may compromise or weaken implementation of the tool and how to deal with them;
- responsibilities in implementing the tool;
- quality criteria and quality control process;
- time schedule;
- resources allocated.

DATA COLLECTION AND ANALYSIS

The evaluation team gathers and analyses all **available documents** (secondary data) that are directly related to the evaluation questions, including:

- management documents – for example, progress reports, steering committee meeting minutes, reviews, audits;
- studies, research papers or evaluations of the evaluand and/or similar evaluands;
- local, national, regional and/or international statistics;
- other relevant documents and other types of information (websites, videos, blogs etc.) available on the internet.

It is clearly not possible for the evaluation team to review all available documents. On the contrary, the team needs to be able to assess the relevance and importance of available information and decide what to focus on in order to answer the evaluation questions.

As part of the desk activities, members of the evaluation team also undertake **interviews** with people who are closely associated with the intervention, including those who have been involved in its design, management and/or supervision. Interviewees should include managers, European Commission (EC) service representatives, and if possible, key partners in the concerned country or countries.

The evaluation questions should not be copied and pasted into interview guides or questionnaires. Evaluation questions are to be answered by the evaluation team, not by the stakeholders. The team can build on stakeholders' statements, but only through careful cross-checking and analysis.

PREPARATION OF DESK REPORT

The team drafts the desk report, which describes all steps taken to date and is delivered at the conclusion of the desk activities. There is no prescribed format for the report. It should be no more than 15 pages, excluding annexes, for most intervention-level evaluations; for other types of evaluations, the length may need to be agreed upon with the reference group during the inception phase.

> **NOTE:** *In evaluations of small interventions, a desk note or slide presentation is often prepared instead of a full desk report.*

Contents

A sample outline for the report is provided in Box 2.4.3. Basically, the report (i) **sums up the findings of the desk activities** in answering the evaluation questions, including the issues still to be addressed and the preliminary hypotheses to be tested during the field activities; and (ii) **outlines the work to be done in the field**, including the people to be interviewed/consulted, the dates and itinerary of the field visits, and assignment of tasks within the team. A further iteration of the evaluation matrix is annexed to the report (see Figure 2.3.3), this time with both Parts A and B completed to the extent possible.

> **OPSYS:** *The evaluation team uploads and submits the draft desk report.*

Review and finalisation

The evaluation manager submits the draft report to the reference group for consultation. If appropriate, the evaluation manager convenes and chairs a meeting at which the report is presented and discussed.

The members of the reference group comment on the draft. The comments are compiled by the evaluation manager and forwarded to the evaluation team. The team then updates the report, in accordance with the following guidance:

BOX 2.4.3 Sample outline of the desk report

1. Introduction
2. Background and key methodological elements, briefly covering:
   - Overall evaluation approach
   - Overview of tools and techniques used
   - Data collection and analyses
   - Challenges and limitations
3. Preliminary answers to each evaluation question, with an indication (in tabular form) of the hypotheses to be tested in the field and information gaps
4. Updated field visit approach and work plan (if relevant; see Box 4.2 for information on intercultural considerations)
5. Main annexes
   - Preliminary answers by judgement criteria
   - Updated evaluation matrix (Parts A and B)
   - List of documentation consulted
   - List of people met with/interviewed

- Preparation of a slide presentation of preliminary findings emerging from the desk activities (free format).
- Remote and/or face-to-face presentation of the preliminary findings from the desk activities to the evaluation manager and the reference group, supported by a slide presentation.
- Revision of the report (as relevant) following receipt of comments and/or slide presentation (or desk note). Comments received should be addressed either in a revised version of the desk report or in subsequent reports.
- Requests for improving methodological quality are satisfied, unless this is not possible, in which case full justification is provided by the evaluation team.
- Comments on the substance of the report are either accepted or rejected. In the later instance, dissenting views are outlined in the report.

The evaluation manager checks that all comments have been properly handled and, if satisfied, approves the report and authorises the launch of field activities.

> **OPSYS:** *The evaluation manager approves the report.*

2.4.2 Field activities

NOTE: *The COVID-19 pandemic underscored that access to the field is not always possible and, in some cases, remote methods of evaluation have to be considered; when structuring your evaluation assess whether field activities are feasible, and to what extent.*

The purpose of the field activities is to conduct primary research and validate/modify the hypotheses formulated during the desk activities. The duration is typically a matter of weeks when carried out by a mission of international experts. The time frame can be extended when local consultants are responsible, with subsequent benefits realised in terms of in-depth investigation and reduced pressure on stakeholders.

PREPARATION

The evaluation manager, supported by the reference group, performs a variety of facilitative and oversight tasks to smooth the evaluation team's data collection and analytic work during the field activities. The respective roles and responsibilities of the key stakeholder groups are summarised below.

Evaluation team

- Prepares the work plan specifying all tasks to be implemented during field activities, including responsibilities, schedules, modes of reporting and quality requirements.
- Provides key stakeholders in the partner country with an indicative list of people to be interviewed, surveys to be undertaken, dates of visit, itinerary and names of responsible team members.

Evaluation manager

- Ensures the work plan is sufficiently flexible to accommodate circumstances in the field.
- Ensures public authorities in the partner country/countries are informed of field missions/visits through the appropriate channels.
- Ensures evaluand managers and key stakeholders are provided with an indicative list of people to be interviewed, dates of visit, itinerary and names of responsible team members.
- Ensures logistics are agreed upon in advance.
- Guarantees adequate contact, consultation with, and involvement of the different stakeholders, including the relevant central and local government authorities and agencies, the final beneficiaries, representatives of civil society and other relevant non-governmental organisations and other donors including European Union Member State agencies.
- Is prepared to interact swiftly and react as quickly as possible at the evaluation team's request if a problem is encountered in the field that cannot be solved with the help of the evaluand manager.

Evaluation manager and reference group

- Facilitate interviews and other data collection methods such as surveys and site visits by appropriate means, including official letters or through informal contacts within the government.
- Facilitate retrieval of any additional documents or data sources and access to key informants in the EC and relevant partner countries for the evaluation team.

Where and as relevant, the evaluation team may hold an information meeting in situ with key stakeholders within the first days of the fieldwork covering the following points:

- presentation and discussion of the work plan;
- access to data and key informants;
- ways to deal with and solve potential problems.

NOTE: *Where multiple countries are involved, the evaluation team will hold a briefing meeting in each visited country, preferably with the participation of the EU delegation.*

DATA COLLECTION AND ANALYSIS

The evaluation team **implements the field data collection plan**. Any difficulties are immediately discussed within the team; where necessary, solutions are discussed with the evaluation manager.

The evaluation team should make use of the **most reliable and appropriate sources of information** (see Box 2.4.4), respecting the rights of individuals to

BOX 2.4.4 Importance of the 'outside' perspective

A key evaluation aim is to determine the extent to which the evaluand's objectives were or are being achieved in terms of both benefits for the targeted group and wider impact. Achievement of objectives is therefore to be judged more on beneficiaries' perceptions of benefit received than on managers' perspective of outputs delivered or results achieved. Consequently, interviews and surveys should focus on outsiders (beneficiaries and other affected groups beyond beneficiaries) as well as insiders (managers, partners, field-level operators). The work plan should clearly state the planned proportion of insiders and outsiders among those to be interviewed/surveyed. Surveying outsiders may require that language and/or cultural gaps be bridged.

provide information in confidence, and being sensitive to the beliefs and customs of local social and cultural environments.

SEE: Section 3.3 *and* Chapter 4 *for further guidance.*

Fieldwork is meant to collect **evidence that is as strong as possible**. This should include:

- direct observation of facts including tangible evidence such as infrastructure (buildings/roads, bridges, irrigation systems, processing plants) etc.;
- statements by informants who have been personally involved in the evaluand;
- indirect reporting on facts by informants who have not been personally involved but have reliable knowledge about the evaluand.

NOTE: *All evaluation team members (and managers) should understand that the evaluation is neither an opinion poll nor an opportunity to express one's preconceptions.*

DEBRIEFING AND REPORTING

One or several debriefing meetings are held at the completion of the field activities to assess the reliability and coverage of data collection, and to discuss the most significant findings. These meetings include all team members; at least one of them is organised with the reference group.

At the meeting(s), the evaluation team presents an overview covering the reliability and scope of the collected data, and the initial analyses and findings. The meeting(s) serves as an opportunity to strengthen the evidence base of the evaluation. No report is submitted in advance, and no minutes are provided afterwards.

NOTE: *In evaluations involving multiple countries, the evaluation team holds a debriefing meeting in each visited country, preferably with the participation of the delegation. A country note is written and circulated to relevant EU delegation stakeholders in the country.*

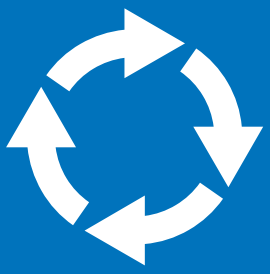
At the conclusion of the field activities, the evaluation team prepares a field note (see Box 2.4.5) in accordance with the specifications set out in the terms of reference (ToR).

NOTE: *In the event that desk and field activities are merged, a single combined desk/field note – an interim report – is prepared. For more information, see the ToR templates and guidance notes available on the EU intranet.*

BOX 2.4.5 Field note

The field note is to be delivered at the end of the field activities. The format is not prescribed but should be no more than 10 pages long, excluding annexes. It should contain at least the following:

- list of activities conducted;
- difficulties encountered and mitigation measures adopted;
- intermediate/preliminary findings;
- preliminary overall conclusions (to be tested with the reference group).



SECTION 2.5

Synthesis phase

The main objective of the synthesis phase is to report on the results of the evaluation. The evaluation team draws up the final report and executive summary, which includes the findings and conclusions that respond to the evaluation questions, as well as an overall assessment of the evaluand, based on robust evidence gathered during the previous phases of the evaluation. The report also includes recommendations, which are clustered and prioritised. The final report is subject to quality assurance (see Section 2.8). The main activities performed in the synthesis phase are:

- analysis and synthesis of the evidence and data collected during the previous phases to provide final, robust answers to the evaluation questions;
- produce an evidence log, if requested in the evaluation terms of reference (ToR);
- preparation and finalisation of a final report and executive summary.

The evaluators ensure that:

- their assessments are objective and balanced, the statements they make are accurate and evidence-based, and the recommendations are realistic and clearly targeted;
- when drafting the report, they acknowledge clearly where changes in the desired direction are already known to be taking place as well as where expected changes are not materialising as planned;
- the writing style of the report takes the users and broader audience into account.

2.5.1 Distilling the findings, conclusions and recommendations

In the synthesis phase, the evaluation team analyses the quantitative and qualitative data gathered and presents these in an easy-to-read and logically structured format (see Box 2.5.1), often organised by the evaluation questions, or in a way that best suits the needs of the target audience. **Conclusions** are then drawn that summarise these findings with a few closing paragraphs, often organised according to the evaluation criteria. **Recommendations** are then prepared to improve or reform the intervention or policy, carefully targeted to the appropriate audiences at all levels, especially within the European Commission (EC) structure. There should be clear links between the findings and the conclusions on the one hand, and between the conclusions and the recommendations on the other (see Figure 2.5.1).

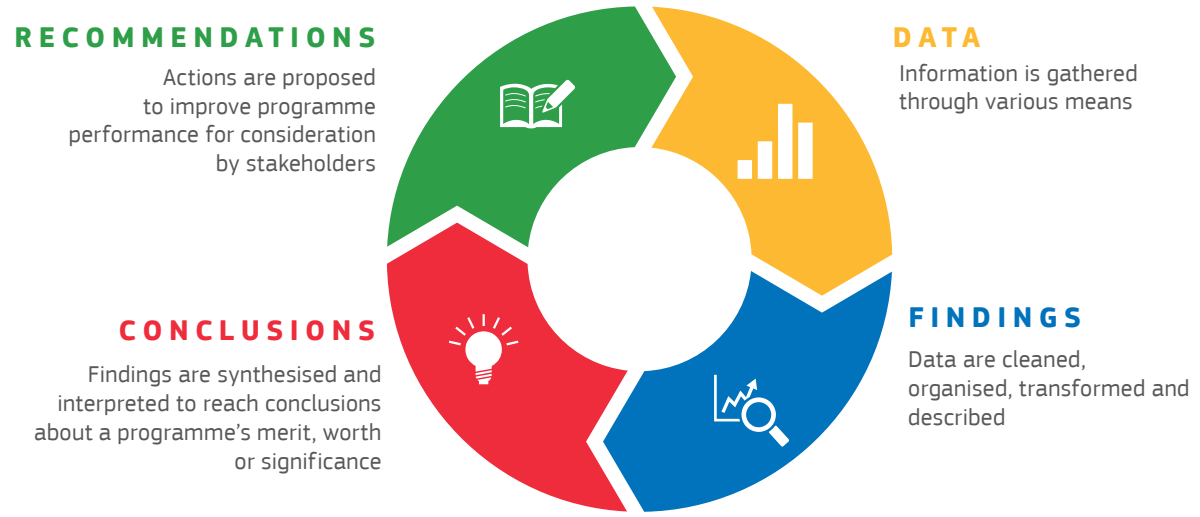
The importance of conclusions and recommendations cannot be overstated. Conclusions are the results of an evaluation study, and provide a summary of the key findings generated from the analysis of data related to the agreed-upon evaluation questions and their judgement criteria. They should in turn lead to actionable recommendations clearly targeted at those responsible for taking that action. Effective evaluation requires that conclusions and recommendations be well founded, closely linked and clearly communicated.

BOX 2.5.1 The importance of clear and tailored communication

Effective communication is essential to ensure that stakeholders understand the evaluation results, the rationale for the recommendations, and their implications for planning and implementation. To communicate evaluation findings and recommendations effectively, the communication approach must be tailored to the audience. For example, programme managers may require a **detailed presentation** of the evaluation results, including the methodology, data analysis and limitations. In contrast, donors may require a **concise summary** of the evaluation results, including the key findings and recommendations. Regardless of the audience, it is essential to use clear, concise and jargon-free language to ensure that **stakeholders understand the evaluation results and recommendations**. Effective communication should also provide opportunities for stakeholders to ask questions, clarify doubts and provide feedback on the evaluation results and recommendations. It should highlight the benefits of the evaluation, such as identifying what worked and what did not as well as opportunities for improvement and lessons learned. This can help build support for future evaluations and ensure that stakeholders understand the value of the evaluation process.

These precepts of good and tailored communication are also critical for effective dissemination, as discussed in Section 2.6.

FIGURE 2.5.1 The evaluation cycle



They should be evidence-based, comprehensive and practical and tailored to the needs of the different stakeholder groups.

Conclusions and recommendations are critical components of any evaluation exercise providing decision-makers with evidence-based information that can be used to **improve the performance of the intervention** being evaluated or to **ensure that future interventions are designed and implemented in a way that maximizes their impact** and achieves their intended objectives. Reviewing the evaluation findings is a crucial step in drawing conclusions and making the corresponding recommendations, as discussed below.

FINDINGS

The evaluation team formalises its findings, which are derived from interpretation and analysis of the evidence (data and information) gathered in the previous phase. This evidence is incorporated into the evidence log (broken down by indicator; see Figure 2.3.3), along with an assessment of the quality of the evidence.

SEE: Section 3.4.

Findings may include cause-and-effect statements (e.g. 'partnerships, as they were managed, generated lasting effects'). But unlike conclusions, findings do not involve value judgements.

The evaluation team proceeds with a detailed review of its findings with a view to confirming or dismissing them. This assessment is done from a critical perspective that requires consideration of the following:

- if **statistical analyses** are used, whether they withstand validity tests;
- if findings arise from a **case study**, whether other case studies contradict them;
- if findings arise from a **survey**, whether they could be affected by a bias built into the survey;
- if findings arise from an **information source**, whether cross-checking indicates contradictions with other sources;
- whether findings could be explained by **external factors** independent of the evaluand;
- whether findings **contradict lessons learned elsewhere** and if so, if there is a plausible explanation for that.

CONCLUSIONS

The evaluation team derives the evaluation's conclusions from the findings and from other issues that have emerged during the evaluation process. The conclusions are generally presented in line with the evaluation criteria.

Conclusions involve value judgements, also called **reasoned assessments** (e.g. 'partnerships were managed in a way that improved sustainability in comparison to the previous approach'). Conclusions are justified in a transparent manner by making the following points explicit:

- Which aspect of the intervention has been assessed?
- Which evaluation criterion was used?
- How was the evaluation criterion applied in this particular instance?

The evaluation team strives to formulate a **limited number of conclusions**. They either clarify or delete any value judgements that are not fully grounded in facts and fully transparent.

The evaluation team synthesises its conclusions into an **overall assessment** of the evaluand, and writes up a summary of all conclusions, which are then prioritised. Methodological limitations are described, as well as any dissenting views.

The evaluation team leader verifies that the conclusions are not systematically biased towards positive or negative views, and checks that any identified weaknesses lead to constructive recommendations.

RECOMMENDATIONS

Recommendations need to be relevant, feasible actionable and likely to lead to improvements in performance. They should address specific issues or

challenges identified during the evaluation, and they should be specific, measurable, achievable, relevant and time-bound (SMART).

Recommendations may include **proposed changes to design or implementation**, such as revising the theory of change, or enhancing the monitoring and evaluation framework. Recommendations can also include **suggestions for future research** to address gaps in knowledge or evaluate the effectiveness of specific components of an intervention.

By being **actionable**, recommendations help ensure that evaluation leads to tangible improvements in performance and to positive change.

LESSONS LEARNED[1]

Identifying lessons learned through evaluation allows programming practices or operational approaches to be highlighted which can then be promoted, avoided or shared with others. Lessons learned should contain knowledge that can be applied to future actions. They should consist of a generalised principle that can be applied in other situations and should be accompanied by an explanation and evidence of why it is considered a lesson learned. A lesson learned should capture a shift in understanding about an activity or process and provide new learning for ongoing or future programming. The context of the lesson needs to be clear, so others can determine its appropriateness and utility to their own situation.

Lessons can be derived from and screened against the following to ensure their value and utility:

- practice wisdom and the experience of practitioners;
- experience from programme participants, clients and intended beneficiaries;
- evaluation findings seeking patterns across programmes;
- basic and applied research;
- expert opinion;
- cross-disciplinary connections and patterns;
- strength of the connection to outcome attainment.

The greater the number of supporting sources for a lesson learned, the more rigorous the supporting evidence, and the greater the triangulation of supporting sources, the more confidence built in the significance and meaningfulness of a lesson learned.

It is important to note that lessons learned are not always about positive outcomes.

2.5.2 Preparing the final report

INITIAL DRAFT

The evaluation team drafts the final report; this draft should be consistent, concise, clear and free of syntactical errors both in the original version and in any translated versions. The draft should be no longer than 40 pages excluding annexes. The use of figures and tables is strongly recommended, as is the use of clear, accessible language.

> **NOTE:** *The required format and contents of the final report are detailed in Annex V of the ToR templates and guidance notes available on the European Union intranet.; Table 2.5.1 describes the report's main sections. If separate country notes were prepared as part of the field activities, these are included as annexes to the draft report.*

The evaluation team also drafts two stand-alone **executive summaries**:

- One version is part of the final report and covers its main points (the evaluation purpose, methods used, main findings, and conclusions and recommendations).
- The second version is prepared for upload to OPSYS and follows a prescribed format.

Before submitting the draft final report, the evaluation team leader checks that it meets the quality criteria as per the quality assessment grid (QAG) and ensures that:

- all the evaluation questions are answered;
- reliability and validity limitations are specified;

[1] This material draws on Patton and Millett (1998).

- the conclusions relate to evaluation criteria in an explicit and transparent manner;
- EU evaluation guidelines have been followed;
- tools and analyses have been implemented according to standards;
- the language, format, illustrations etc. are according to the standards set out.

> **OPSYS:** *The evaluation team uploads and submits the draft final report and executive summary. The online executive summary is submitted using the specific PDF or web form.*

REVIEW AND QUALITY ASSURANCE

The evaluation manager receives the draft final report and assesses its quality using the prescribed QAG, as described in Subsection 2.8.3. The quality assessment should enhance the credibility of the evaluation without undermining its independence. It focuses on the way conclusions are substantiated and explained and not on the substance of the conclusions.

The evaluation manager submits the draft report to the **reference group members for consultation**. If appropriate, the evaluation manager convenes and chairs a meeting where the report is presented and discussed. Special attention is paid to the utility of conclusions and feasibility of recommendations.

FINALISING THE REPORT

The evaluation team leader receives the quality assessment of the draft final report from the evaluation manager, including the QAG, and prepares a revised draft of the final report, executive summary and annexes based on the comments received.

> **NOTE:** *It is recommended that the evaluation team prepare a summary table listing all the comments received, and specifying the modifications made to address these or explaining the reasons for not taking action on a particular comment. This response should then be attached to the QAG and the next iteration of the report.*

The evaluation team then presents the revised report in a reference group meeting. The presentation is supported by a series of slides which cover:

- answered questions and methodological limitations;
- overall assessment, conclusions and lessons learned;
- recommendations.

> **OPSYS:** *The evaluation team uploads and submits the final report and executive summary. The executive summary must be submitted online using the specific PDF or web form.*

The evaluation manager checks that the comments received have been taken into account appropriately, and that the report is ready for dissemination, including the full set of annexes.

The evaluation manager approves the final version of the report and carries out a final quality assessment using the QAG, writing qualitative comments for all criteria, and deciding upon the overall quality rating, which is sent to the reference group members and the contractor.

> **OPSYS:** *The evaluation manager approves the final report and executive summary; this includes approving all dissemination products to be prepared and distributed in Section 2.6.*

The report is printed out according to the instructions stated in the ToR.

> **OPSYS:** *OPSYS will save the digital version of the final report, the executive summary as well as all the annexes and other documents (e.g. slides presented by the evaluation team or minutes from meetings).*

TABLE 2.5.1 Contents of the final report

Executive summary	Highlights the evaluation purpose, the methods used, the main evaluation findings and the conclusions and recommendations. Should be a stand-alone document.
1. Introduction	Presents a description of the evaluand, of the relevant country/region/sector background and of the evaluation, providing sufficient methodological explanations to establish the credibility of the conclusions and to acknowledge limitations or weaknesses, where relevant.
2. Findings	Presents the answers to the evaluation question headings, supported by evidence and reasoning. Findings by judgement criteria and detailed evidence by indicator are included in an annex to the report.
3. Overall assessment (optional)	Synthesises all answers to the evaluation questions into an overall assessment of the evaluand. The aim is to articulate all the findings, conclusions and lessons in a way that reflects their importance and facilitates readability. The structure should not follow the evaluation questions, the logical framework or the evaluation criteria.
4. Conclusions and recommendations	**4.1 Conclusions.** These are organised by evaluation criterion. In order to allow better communication of the evaluation messages that are addressed to the Commission, a table organising the conclusions by order of importance can be presented, or a paragraph or subchapter emphasising the three or four major conclusions presented in order of importance, while avoiding repetition.
	4.2 Recommendations. These are intended to improve or reform the intervention in the framework of the cycle underway, or to prepare the design of a new intervention for the next cycle. Recommendations must be clustered and prioritised, and carefully targeted to the appropriate audiences at all levels, especially within the Commission structure.
	4.3 Lessons learned. Generalises findings and translates past experience into relevant knowledge that should support decision-making, improve performance and promote the achievement of better results. Ideally, they should support the work of both the relevant European and partner institutions.
5. Annexes	• Terms of reference of the evaluation • List of activities specifically assessed • Names of the evaluators; curricula vitae (CVs) can be included, but should be summarised and limited to one page per person • Detailed evaluation methodology, including the evaluation matrix, options taken, difficulties encountered and limitations, details of tools and analyses • Detailed answer by judgement criteria • Evaluation matrix with data gathered and analysed by evaluation question and judgement criteria indicators • Intervention logic/logical framework matrices (planned/real and improved/updated) • Relevant geographic map(s) where the intervention took place • List of people/organisations consulted • Literature and documentation consulted • Other technical annexes as relevant (e.g. statistical analyses, matrix of evidence, databases, list of documents used)

SECTION 2.6

Dissemination phase

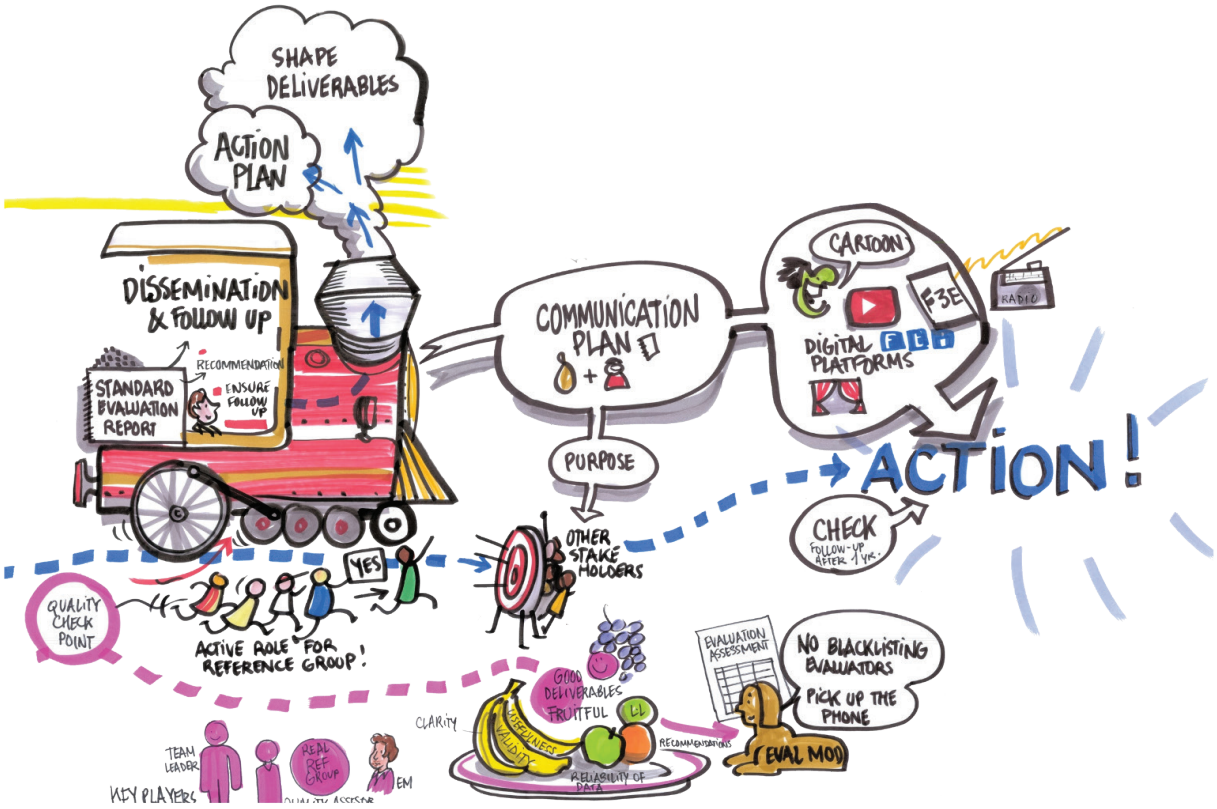
During the dissemination phase (Figure 2.6.1), the evaluation team prepares **knowledge products** (videos, infographics, fact sheets, podcasts, seminars etc.) to disseminate the evaluation results, as set out in the terms of reference (ToR) and/or inception note/report. This includes disseminating the final report itself.

2.6.1 Disseminating the evaluation report

Finalising the evaluation report ends the evaluation process per se, but sets off a productive new period of learning and feedback – which is launched by disseminating the evaluation report and findings widely, wisely and well. Getting the report and its findings into the right hands in a timely and appropriate manner is key:

- It publicises the evaluation's conclusions, lessons learned and recommendations to promote transparency and use both within the European Commission (EC) and across European institutions, external partners, networks of experts, the media and the wider public.
- It should stimulate discussion and help with the identification of appropriate follow-up actions to put into practice the lessons learned and feed the evaluation findings into the next cycle of decision-making and programming.
- It offers the chance to close any existing feedback gaps with evaluation respondents, and the wider stakeholder group, which can ensure better uptake and use of the learning and knowledge produced.

FIGURE 2.6.1 **Dissemination at a glance**



SEE: Disseminating Evaluations *on Capacity4Dev's Evaluation methodological approach wiki for further discussion of the importance of dissemination within the EC.*

2.6.2 Thinking about dissemination

All too often, disseminating evaluation findings is limited to emailing a web link for the final report and executive summary to the people who commissioned the evaluation and the people who participated in it. That is far from adequate and rarely achieves the goals and objectives of disseminating the findings, conclusions and learning to be distilled from the report.

There are three key interrelated questions to be answered about dissemination of information:

- What should be shared?
- With whom should it be shared?
- How should it be shared?

NOTE: *The answers to these questions can be captured in a matrix that spells out who to tell what in an appropriate manner for maximum impact. Once the key messages for different targeted audiences have been determined and the dissemination product options identified, the next step is to check if those options are feasible given the available financial and human resources, and the time available.*

Following is general guidance in preparing to spread the knowledge from the evaluation.

Plan for dissemination from the beginning of the evaluation. A dissemination plan should be specified in the ToR – including what dissemination is to be carried out and who will be in charge of developing the different dissemination products and for what audiences – and budgeted for in the evaluation contract.

Determine the audience. Decide on the target audience: internal and external – who will benefit most from the evaluation findings? Thinking about the type of audience(s) to reach will help in defining the type of message to convey, and the style as well as the method of dissemination.

Decide on key messages to share with the audience. The method of presentation will impose limits on the amount of information that can be conveyed. For example, focusing on the top 5 or 10 highlights of the evaluation findings will best suit most visual/spoken/interactive formats (see Table 2.6.1). For an easy tip, answer this question: If the audience can take away only one or two key messages, what should these be?

Determine the appropriate presentation/format. The uptake of evaluation results is often hampered by the way evaluation reports are presented. Dissemination of the final report and the executive summary is generally a standard requirement within the Directorate-General for International Partnerships (DG INTPA) and the Service for Foreign Policy Instruments (FPI) (see Box 2.6.1), but these documents are not always suitable for non-specialist readers. More appropriate formats can be used by focusing on specific user groups and purposes (see Table 2.6.1). Be innovative – an evaluation that only ends up on a shelf is a waste of resources. Some alternative ways to disseminate evaluations are:

- newsletters;
- briefs and fact sheets;
- seminars and validation events;
- webinars and online presentations;
- blog posts and interactive discussions via staff intranets, communities of practice or knowledge networks;
- infographics;
- multimedia/video presentations;
- podcasts;
- photo stories;

> **NOTE:** *Use photos and videos with care and with the informed consent of the subjects pictured, keeping in mind their protection and dignity. For more information, see OECD DAC (2022).*

- summary findings table with a simple rating system highlighting strengths and weaknesses;
- scorecards or dashboards with key data, quotes and findings;
- interactive web pages including maps.

> **NOTE:** *Examples of these different products and channels are available on* Disseminating

TABLE 2.6.1 Dissemination knowledge product options by communication mode

Type	Description	Example
Visual	Using a visual format is an excellent way to communicate a lot of information in a small space. Some visual presentations can also be included in written reports, summaries, blogs or presentations.	• Infographics • Illustrations and cartoons • Data dashboards • Posters • Photographs
Spoken	Using a spoken format presentation provides opportunities to hear the actual voices of evaluation participants. It is also an engaging and interactive way to communicate findings.	• Evaluation dissemination seminars • Presentations • Podcasts • Videos • Music, spoken word (or even interpretive dance)
Written	In addition to traditional evaluation reports, consider using shorter, more action-oriented formats.	• Summary reports • Blogs • Newsletters • Postcards

SOURCE: Hassnain, Kelly and Somma (2021).

BOX 2.6.1 DG INTPA and FPI dissemination requirements

- Fifteen days (on average) after approval of the final report, if requested by management, the evaluation manager should publish the final report, the executive summary and the quality assessment grid (QAG) on the EC intranet.
- The executive summary can also be posted on the relevant service intranet sites.
- The evaluation manager should draw up the dissemination list and send the final evaluation report and/or the executive summary to the relevant services and to the partners.
- A short article can be written to facilitate dissemination of the main conclusions and recommendations.
- The evaluation manager should prepare a fiche that provides a summary of the recommendations and requests opinions from the services to which the recommendations are addressed. (See Section 2.7.)
- The responses from the different services should be published on the Commission's website.

Evaluations *and* Disseminating knowledge generated by evaluations. Some examples from the EU. *on Capacity4Dev's Evaluation methodological approach wiki.*

2.6.3 Disseminating the final report

Box 2.6.1 sets out the minimum requirements of evaluation dissemination, focusing on the final evaluation report and the executive summary. Dissemination of these products could be to:

- senior management of DG INTPA, FPI, the European Union (EU) delegations and regional teams;
- other EU institutions;
- representatives of the EU Member States;
- partner countries;
- thematic/sector experts and their groups;
- evaluation participants;
- other stakeholders.

To distribute the report/executive summary, the evaluation manager circulates the full-length report to the relevant Commission services and other evaluation users. This dissemination could be followed up with one or several presentations targeted at specific audiences, such as expert networks in the country or region, the media, government-donor coordination bodies and/or non-state actors. The evaluation team may be asked to participate in these presentations.

OPSYS: *To the extent that the OPSYS platform can accommodate, the evaluation manager should upload all dissemination products from the evaluation.*

2.6.4 Disseminating beyond the final report

This subsection describes ways of disseminating evaluation knowledge beyond directly sharing the evaluation report with all relevant stakeholders. These are presented in no particular order, as there is no preferred technique, but rather what works best for the specific target audience.

Supplementary publications. The full evaluation report and executive summary can be supplemented and complemented by the following:

- A one-page summary can be written specifically for the service that managed the evaluation, highlighting the main conclusions and recommendations.
- A summary aimed at the international development community can highlight transferable key lessons learned; this can be posted to the web with a link to the full report.
- One or more articles may be written for the general public or specialised networks.

Dissemination seminars. Whether online or in person, dissemination seminars are a particularly effective means of disseminating evaluation

results to interested audiences. They provide a good opportunity to share key messages with stakeholders and to highlight the most important lessons emerging from the evaluation with a view to increasing ownership of the evaluation findings and use. Seminars also help increase visibility about the evaluand. Furthermore, they can be an occasion to stimulate debate on specific issues covered in the evaluation so as to provide additional inputs to the EC, national policymakers and operational staff.

Webinars. Evaluators can be asked to present their findings through a webinar to a wider global audience via the use of social media to promote key messages and knowledge products.

> **SEE:** How-to Guide on Evaluation Dissemination Seminars *available for download from Capacity4Dev's Evaluation methodological approach wiki. Also see Hassnain and Lorenzoni (2020a)*

Thematic discussions. Wherever feasible, and where there is new and useful knowledge generated by and on the subject of an evaluation, it is recommended to organise thematic discussions targeting relevant stakeholders. These discussions could be organised online or in person depending on the target audience.

Validation workshops. For complex interventions and in complex settings, a validation workshop could be conducted at the community level to communicate evaluation findings to participants in the evaluation (particularly in the most marginalised communities) to close the feedback loop. Besides the commonly practiced wrap-up meetings in capital cities targeted at public figures/leaders, it is suggested that evaluators and/or intervention implementers share evaluation findings with the people who benefited from the intervention (or otherwise) and who will support the scaling up of that learning for deeper and long-lasting impact on the ground (Hassnain, 2021).

Other options for dissemination can contribute to the global knowledge base, such as the EC Joint Research Centre (JRC), participating agencies' websites, government agencies involved, the Global Evaluation Initiative etc.

SECTION 2.7

Follow-up phase

In the follow-up phase, the evaluation manager takes steps to follow up on the recommendations proposed by the evaluation – first by ensuring that the responsible stakeholders receive all relevant evaluation deliverables and second by following up with them on these a year later.

An evaluation's true value is evidenced after it concludes. What changes does it prompt? How are interventions – and thus ultimately people's lives – improved? An evaluation is a sterile accounting exercise without follow-up. That follow-up begins when the knowledge arising from the evaluation is shared via traditional and innovative dissemination means and methods, as discussed in the previous section. It is spurred by actions taken by the evaluation manager.

After the evaluation report and dissemination knowledge products are circulated to relevant stakeholders, the evaluation manager draws up a list of the evaluation's recommendations as captured in the executive summary of the final report. Using a template in OPSYS, the evaluation manager identifies, for each recommendation, the relevant service/stakeholder to take action on the recommendation.

The evaluation manager then solicits written feedback from each relevant service/stakeholder about each recommendation, specifically:

- its importance/priority (high, medium or low; short, medium or long term);
- the extent of agreement with the recommendation (yes, no or partially);

- the justification for those decisions about priority and agreement and the actions to be taken to address the recommendation.

The evaluation manager collects all comments received and records the responses, noting:

- the service's extent of agreement with / acceptance of the recommendation;
- the name of the person in charge of implementing the recommendation;
- the planned date of completion;
- any comments.

> **NOTE:** *In addition to capturing this information on OPSYS, the evaluation manager can elect to publish service responses to the evaluation on the same web page as the final report.*

Approximately one year later, the evaluation manager follows up with the services to determine the extent to which they acted on the tasks planned for each recommendation. The evaluation manager again solicits written feedback from each relevant service/ stakeholder:

- indicating if all accepted recommendations have been implemented 6–12 months later;
- describing the actions that have been taken, what the results have been and the main problems encountered.

> **OPSYS:** *The evaluation manager fills in and confirms as final the one-year follow-up response.*

A completed evaluation and its conclusions and recommendations should feed into subsequent stages of planning and intervention design and help address challenges by providing more evidence on the validity (or not) of the theories of change and data for comparison and reference: this is the basis of the 'evaluate first' principle.

Where feasible, the evaluation results should feed into the annual activity reports, and related follow-up actions should be identified in the annual management plans of the Commission services. Identifying and sharing planned follow-up actions is part of accepting responsibility and accountability for EU actions and ensures transparency.

Design documents such as the Action Document template specifically require an explanation of how the proposed intervention will build on lessons learned from previous, similar actions and how these have been incorporated into the design. The following guidance is cited in Section 3.4 of the Action Document template:

> Lessons learnt are the outcomes of a learning process, which involves reflecting upon the experience. The key questions to be answered are: (1) What has and has not worked in the past? Lessons learnt can be horizontal or sectorial. (2) Which were the enabling and limiting factors? (3) How are these lessons considered in the current action? Which stakeholders will act upon them?

In many instances, the immediate dissemination/ follow-up to an evaluation is identified in the legal base of the intervention and takes the form of a Commission report to the European Parliament and the Council on the findings of the evaluation.

The final report, along with the communication and dissemination products discussed in Section 2.6 could be examined periodically through meta-analysis to further build a national, regional and global evidence base. Follow-up events on the key action points of the management response would also help ensure better utilisation and uptake of the evaluation findings.

SECTION 2.8

Quality assurance

All evaluation outputs produced in the various phases of the evaluation must meet quality standards. Therefore quality assurance is not considered a specific phase of an evaluation, but runs throughout the entire evaluation process from the drafting of the terms of reference (ToR) through to the dissemination of findings/results and follow-up.

The evaluation manager begins by identifying the key players involved in quality assurance and clearly defines the key steps in quality assurance across the evaluation process.

The importance of maintaining quality across the phases and with reference to each output/deliverable means that problems can be resolved well before the final report is submitted and it is too late to remedy a quality problem that had existed from an earlier stage.

2.8.1 Roles and responsibilities

EVALUATION MANAGER

The evaluation manager is responsible for ensuring the quality of the evaluation by:

- ensuring that the selected evaluation team has sufficient expertise in conducting evaluations – in particular, the evaluation skills of the team leader are crucial to ensuring high-quality evaluation outputs;
- establishing quality checkpoints at different phases in the evaluation process;

- mobilising the reference group to obtain feedback on quality;
- defining rules that deal with quality problems.

The evaluation manager holds ultimate responsibility for ensuring the methodological quality of all evaluation deliverables. This further entails:

- resisting any temptation to 'negotiate' the contents of the final report;
- respecting the evaluators' opinions;
- ensuring that quality issues are addressed in a timely manner by the evaluation team leader or the contractor;
- ensuring at an early stage that the reference group members accept criticism.

REFERENCE GROUP

The reference group supports the evaluation manager in ensuring the quality of the evaluation process and products. In addition to the support the reference group members provide in the early stages in setting up the evaluation, they also receive all draft reports and outputs and provide feedback with a view to ensuring the highest level of quality.

EVALUATION TEAM LEADER

The leader of the evaluation team is a key player in ensuring the quality of the evaluation process and products. The evaluation team leader is responsible for checking the quality of data and analyses against the quality criteria set for each evaluation tool and against general principles such as:

- clear and credible presentation of the evaluation findings;
- clear presentation of the applied methodology;
- clear description of any limitations encountered;
- transparent assessment of the biases and reliability of the data underpinning findings;
- confirmation of compliance with the work plan and/or justification for any adjustments;
- confirmation of compliance with anonymity and other ethical rules.

CONTRACTOR/EVALUATION TEAM

All framework contractors are responsible for the quality of the process, the evaluation design, and the inputs and the outputs of the evaluation, as specified under Article 6 of the Global Terms of Reference and the Global Organisation and Methodology of the *framework contract SEA 2023*. The **contractor**:

- supports the team leader in his/her role, mainly from a team management perspective – in this regard, the contractor should make sure that, for each evaluation phase, specific tasks and outputs for each team member are clearly defined and understood;
- provides backstopping and quality control for the evaluation team's work throughout the assignment;
- ensures that the evaluators are adequately resourced to perform all required tasks within the time frame of the contract.

> **NOTE:** *It is strongly recommended that, at the outset of the evaluation, the evaluation manager and the contractor agree on a clear roadmap or plan that explains how quality issues will be addressed if and when they arise. It is also strongly recommended that sufficient time for quality assurance is factored into evaluation planning. In general, it is very unlikely that the first drafts of evaluation deliverables will meet quality standards, so the time required for feedback from the evaluation manager and the reference group, as well as for evaluation teams and contractors to address quality issues, has to be planned for from the outset. Although this will extend the timeline (and cost) of evaluations, it is a very worthwhile investment.*

The **evaluation team leader** prevents major risks threatening quality and ensures that each report/output undergoes a detailed quality check.

The **quality assessor(s)** designated by the contractor carefully checks each evaluation output for quality.

2.8.2 Key steps in quality assurance

Managing the quality of an evaluation starts from the outset (see Figure 2.8.1). Because the ToR forms the

foundation of an evaluation, it is essential that the ToR is well drafted with clearly defined objectives, scope, indicative evaluation questions, methodology and planned deliverables, including dissemination products. If the ToR is weak, the entire evaluation process and resulting outputs will be too.

To ensure the quality and relevance of the final evaluation outputs, the evaluation manager needs to put in place quality checks throughout the evaluation process – that is, to gradually construct quality and avoid discovering a problem in the final stages.

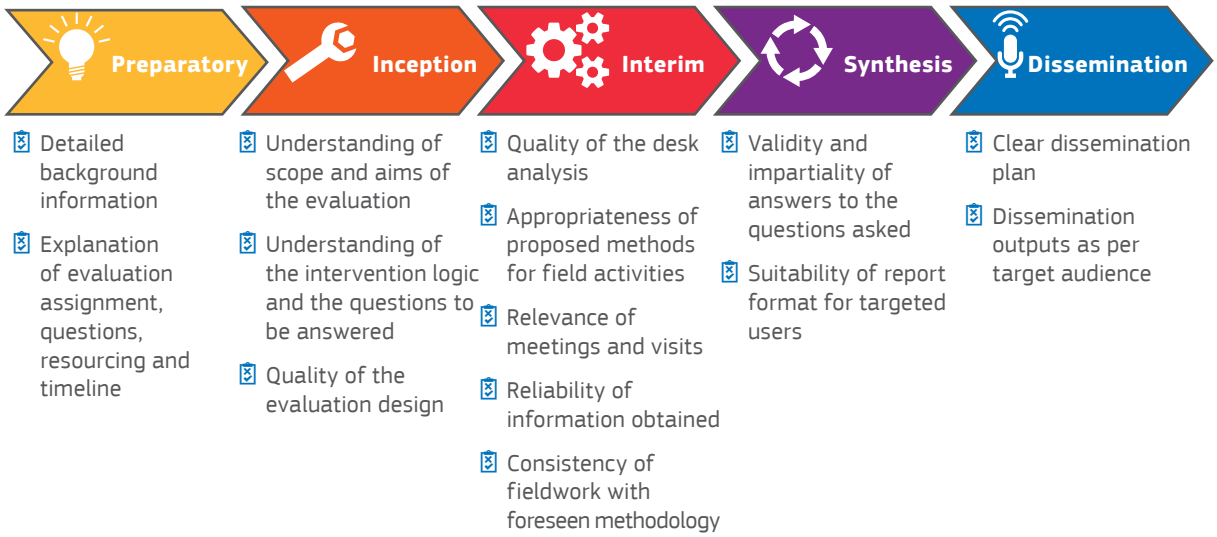
2.8.3 The quality assessment grid

The quality of the draft and final versions of the final report and the executive summary are assessed by the evaluation manager against six separate criteria using the online quality assessment grid (QAG) in OPSYS. The assessment is double checked by a second person who serves as a backup to the evaluation manager.

> **OPSYS:** *The evaluation manager prepares the draft QAG.*

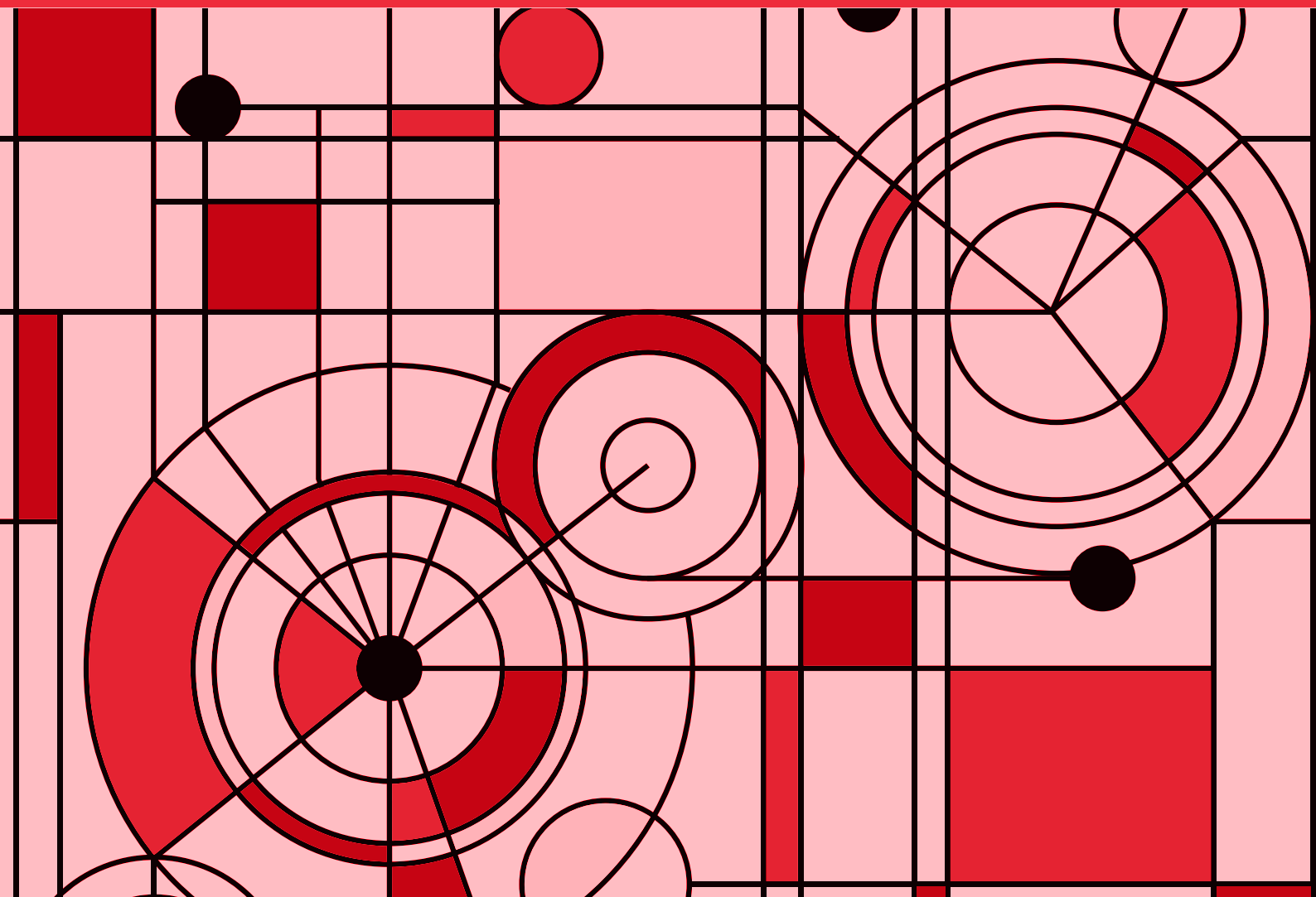
The contractor is given the opportunity to comment on the assessments formulated by the evaluation manager through OPSYS. The QAG will then be reviewed, following the submission of the final versions of the final report and executive summary.

FIGURE 2.8.1 Key quality assurance checkpoints along the evaluation process



3

Approaches, methods and tools



What is this chapter about?

This chapter aims to provide information for readers seeking more detail on topics mentioned elsewhere in the handbook such as evaluation criteria, questions, design, approaches and methods as well as data collection tools, management and analysis.

How will this help you in your work?

This chapter will help you to better understand the evaluation process by providing detailed information on different evaluation approaches, methods and tools. It is not meant to be read in its entirety but rather to serve as a reference for more details on specific aspects of evaluations.

For definitions of key terms used in this handbook, refer to the glossary.

SECTION 3.1

Evaluability, evaluation criteria and evaluation questions

This section looks at the issue of evaluability and how to conduct an evaluability assessment. It then addresses in some detail the six Organisation for Economic Co-operation and Development Development Assistance Committee (OECD DAC) evaluation criteria of relevance, coherence, effectiveness, efficiency, impact, and sustainability, as well as the European Union (EU)–specific one of EU added value. The rest of the section is dedicated to evaluation questions; how to formulate them and identify corresponding judgement criteria and indicators that will form the basis of the evaluation matrix.

SEE: *Discussion of the evaluation matrix in Subsection 2.3.5.*

3.1.1 Evaluability

Among international development agencies there is widespread agreement on the meaning of the term 'evaluability'. The following definition from OECD DAC is widely quoted:

> The extent to which an activity or project can be evaluated in a reliable and credible fashion (OECD-DAC, 2010; p. 21).

This determination is made through a systematic **evaluability assessment**. This is a formal process to inform the timing of an evaluation and improve the prospects of an evaluation's producing useful results. If done properly, an evaluability assessment can save significant resources in terms of time, money and personnel and/or clarify the maximum benefit of an evaluation to its potential users in a given context.

PURPOSE OF AN EVALUABILITY ASSESSMENT

An evaluability assessment asks specifically:

- **whether it is plausible to expect change as depicted in the results chain** – for example, whether there are logical and evidence-informed links between activities, outputs, outcomes and longer-term impacts;
- **whether it is likely to be feasible to evaluate** – are the planned results SMART (specific, measurable, attainable, relevant and time-bound), what data are available, what time and other resources would be needed to fill important gaps in the available data, and are baseline data available against which change can be measured;
- **whether it is likely to be useful to evaluate** – in particular, whether there are specific intended uses for the evaluation.

An evaluability assessment will usually serve multiple purposes, including:

- assessing the overall **feasibility** of evaluation, including appropriate timing;
- improving **allocation** of scarce evaluation resources (people, time, budget etc.);
- gaining a clearer picture of the intervention and its **objectives**, sometimes leading to a revision of its design, including the contents of the intervention logic;
- identifying which **data** are necessary/available and how they can be obtained;
- scoping the interest of **stakeholders** as to their participation in and use of the evaluation;
- designing a feasible **methodology** for evaluation;
- making decisions on evaluation **priorities**.

CONDUCTING AN EVALUABILITY ASSESSMENT

Table 3.1.1 provides a checklist of questions to guide an evaluability assessment. The typical attributes of an intervention that make it evaluable are as follows:

- The expected results are specific, clear and plausible.
- Indicators are available that can be supplied with robust data to assess the results.
- Baseline data and contextual information are available against which change can be measured.
- Relevant risks are identified in a systematic manner and mitigation measures proposed.

SEE: Subsection 3.2.3 *for a review of the components of sound design.*

If an evaluability assessment finds that an intervention is **not ready** for evaluation, then further work might be needed to prepare it – for example, developing an agreed-upon intervention logic or clarifying intended uses.

SEE: *Austrian Development Agency (2022), Davies (2013) and Trevisan and Walser (2014) for more information about evaluation assessment.*

3.1.2 Evaluation criteria

As noted in Subsection 2.2.4, the six evaluation criteria established by OECD DAC are a de facto standard in evaluation worldwide, capturing key aspects of a strategy, policy, instrument, modality, intervention or group of interventions. Within the EU, these evaluation criteria – relevance, coherence, effectiveness, efficiency, impact and sustainability – are joined by a seventh EU-specific evaluation criterion: EU added value.

OECD DAC highlights that the evaluation criteria should be understood within the broader context of two principles:

1. The criteria should be applied thoughtfully to support high-quality, useful evaluation.
2. The use of the criteria depends on the purpose of the evaluation. The criteria should not be applied mechanistically.

SOURCE: Evaluation Criteria *web page on the OECD website.*

TABLE 3.1.1 Questions to ask about the evaluability of an intervention design

Clarity	• Are the long-term impact and outcomes clearly identified? • Are the proposed steps towards achieving these clearly defined?
Relevance	• To what extent can the intervention be considered relevant to the issue(s) it aims to address?
Plausibility	• Is there a robust results chain depicting the logical and evidence-informed links between activities, outputs, outcomes and longer-term impacts? • Is it likely, based on the intervention plan, that the intervention's objective could be achieved within its lifespan? Is there evidence from elsewhere that it could be achieved?
Validity and reliability	• Are there valid indicators (output, outcome and impact levels) for each expected result – i.e. will they capture what is expected to happen? • Are the indicators reliable – i.e. will observations by different observers find the same thing?
Testability	• Is it possible to identify which linkages in the results chain will be most critical to the success of the intervention, and thus should be the focus of evaluation questions?
Contextualisation	• Have assumptions about the roles of other actors outside the intervention been made explicit (both enablers and constrainers)? • Are there plausible plans to monitor these in any practicable way?
Consistency	• Is there consistency in the way the intervention logic is described across various intervention documents (design, monitoring and evaluation plans, work plans, progress reports etc.)?
Complexity	• Are multiple interactions expected between different intervention components, complicating attribution of causes and the identification of effects? • How clearly defined are the expected interactions?
Agreement	• To what extent do different stakeholders hold different views about the intervention objectives and how they will be achieved? • How visible are the stakeholders who might be expected to have different views?

SOURCE: Based on Davies (2013).

OVERVIEW OF THE CRITERIA

Concise definitions of the criteria are given below; fuller definitions are available on the OECD Evaluation Criteria web page.

- **Relevance.** Is the evaluand doing the right things?
- **Coherence.** How well does the evaluand fit with other interventions?
- **Effectiveness.** Is the evaluand achieving its objectives?
- **Efficiency.** How well are resources being used?
- **Impact.** What difference does the evaluand make?
- **Sustainability.** Will the benefits last?
- **EU added value.** To what extent does the intervention bring additional benefits compared to what would have resulted from Member States' interventions only in the partner country? Evaluators are thereby asked to verify whether Member States alone could have resolved the identified problems sufficiently and whether the EU had the competence to act (i.e. a legal basis), and was best placed to do so. EU action should be necessary and should deliver added value compared to the actions of the Member States at central, regional or local levels.

NOTE: *This criterion stems directly from the principle of subsidiarity defined in Article 5 of the Treaty on European Union.*

The specific scope of an evaluation may suggest that it is not necessary to cover all the OECD DAC criteria, but this needs to be justified, as noted by Tool #47 in the Better Regulation Toolbox (EC, 2023). For instance, in some cases during a midterm evaluation, it may be premature to assess impact and/or sustainability;

conversely, during an ex post evaluation, it may be too late to assess efficiency and more cost-effective to focus the attention of evaluators on impact and sustainability.

Alternatively, the scope of a particular evaluation may call for addressing additional or alternative criteria in line with the thematic areas tackled by the evaluand – for example, gender mainstreaming, environmental sustainability and inclusion. Evaluators may also be asked to reflect on specific issues such as conflict sensitivity, coordination or visibility. The point is that the evaluation focus should respond to clearly defined needs, and this focus will be captured by and reflected in the evaluation questions.

THE EVALUATION CRITERIA AND GENDER-RESPONSIVE EVALUATION

To ensure that evaluations are gender responsive, the seven evaluation criteria can include a specific gender lens as described below.

Relevance

- How was gender analysis of the context, sector, problem and stakeholders considered during the formulation of the intervention and/or reformulation in case of changes during implementation? Was any analysis done of how inequality on the grounds of gender intersect with different inequalities or discrimination on the basis, for instance, of ethnicity, age, sexual orientation, social group etc.? How was gender equality integrated in the intervention logic?
- Was the process of consultation leading to the formulation of the intervention purposefully inclusive of stakeholders such as relevant civil society organisations working on gender equality and women's empowerment and aimed at empowering marginalised/excluded groups to address obstacles and promote human rights? Were women and men from a range of diverse social groups, ages and abilities represented? Were intersectional perspectives taken into account? What measures were taken to guarantee meaningful participation of stakeholders (e.g. timely notification, language, location and timing?)
- Was gender equality taken into account and included throughout the intervention (design, implementation and monitoring)? How was it done? If not, why not? Was it consistent with national policies or international instruments on gender equality and relevant international human rights obligations? How? If not, why not?
- To what extent did the planned activities address the causes of gender inequality and discrimination and reach the relevant beneficiaries, including those who are marginalised or disadvantaged?
- To what extent has the intervention effectively contributed to the creation of favourable conditions for advancing gender equality?

Coherence

- To what extent have the results of the intervention complemented/been supported by other EU interventions in the area of external action and foreign policy?
- To what extent was the intervention coherent with EU commitments and strategies in the area of gender mainstreaming/gender equality and with EU Member States' action throughout its life?
- To what extent did it contribute to the implementation of the EU Gender Action Plan and other regional policy documents that include references to gender equality?
- To what extent have the results of the intervention complemented/been supported by the human rights components/interventions of individual EU Member States?
- To what extent was the intervention coherent with those of other donors throughout the programming period in the area of gender mainstreaming/gender equality?

Efficiency

- Were resources (financial, time, people, technical and gender expertise) sufficient to address the gender inequalities defined during the formulation of the intervention? Were they spent or allocated to target the structural causes of inequality? Were these resources easily and unambiguously identifiable? Were these resources consistently allocated throughout and over time? If they were

not consistently allocated, what are / will be the costs of not doing so?

- Were gender- and age-specific constraints taken into consideration when implementing activities? Did the internal monitoring system integrate and use gender analysis and, if so, in what ways?
- What outputs have been received respectively by men and women, boys and girls as a result of the intervention?

Effectiveness

- To what extent did the intervention contribute to achieving its expected results, respectively for men/boys and women/girls, and for those marginalised or in a vulnerable situation? What expected and unexpected results were achieved for women and girls, and for men and boys, also taking into account an intersectional perspective, where relevant? Who benefited most, how and why? What factors played in favour or against the achievement of the expected results, respectively for men/boys and women/girls?
- Has management of the intervention taken care of its gender mainstreaming/gender equality objectives within the wider context of a rights-based approach and translated those objectives into specific actions? How has this been done?
- Were specific risks and challenges inherent to the achievement of gender mainstreaming/gender equality adequately taken into consideration and mitigated? How? What assumptions were made with regard to the gender division of rights, labour, responsibilities etc.? Were these assumptions accurate and relevant?
- Do the results validate the intervention logic in the area of gender mainstreaming/gender equality? How so?

Impact

- What specific impact contributions did the intervention logic foresee for the intervention in the area of gender mainstreaming/gender equality?
- To what extent did the intervention understand and address the underlying causes of gender inequality?
- What is the likelihood that the intervention will have expected/unexpected impacts on human rights and gender mainstreaming/gender equality? Are they expected to be positive or negative, and in which ways will they affect the different stakeholders?

Sustainability

- Did the intervention promote sustainable changes in the area of gender mainstreaming/gender equality? How? What more could have been done to promote greater sustainability with regard to gender mainstreaming/gender equality and changes in gender power relations? If so, how?
- Was an appropriate exit strategy planned for and implemented? How did this strategy address elements of gender mainstreaming/gender equality? To what extent and how were the local partners and different beneficiaries (including rights holders and duty bearers) involved in defining and implementing the exit strategy?
- To what extent do the partners of the intervention own its results in the area of gender mainstreaming/gender equality, and to what extent are they committed to their sustainability after the end of the intervention?

EU added value

- To what extent does the intervention add benefits to or link to Member States' interventions in the area of gender mainstreaming/gender equality?
- To what extent can the results of the intervention in the area of gender mainstreaming/gender equality trigger further bilateral interventions by the EU Member States?

3.1.3 Evaluation questions

Evaluation questions are the **backbone of an evaluation**. They define what the evaluation should concentrate on, have a primary impact on the methodology the evaluation team develops, and determine the findings that will be produced by the evaluation. Evaluation questions give focus to the evaluation and ensure that the evaluation team

emphasises points of primary, rather than secondary, interest.

> **SEE:** *Tool #47 in the Better Regulation Toolbox (EC, 2023); the How-to Guide on evaluation questions in the Evaluation wiki; and Evaluation Questions Checklists on Rick Davies's Monitoring and Evaluation NEWS news service for more detailed information on formulating evaluation questions.*

ORGANISATION

Evaluation questions can be organised in various ways.

- **By the selected evaluation criteria.** In this case, each selected evaluation criterion should be covered by at least one evaluation question (see Table 3.1.2).
- **By clusters, covering transversal areas.** Examples of such areas are (i) policy framework and responsiveness, (ii) management and governance (institutional set-up), (iii) EU cooperation potential (Team Europe approach) and EU added value, (iv) partnerships (engagement, coordination and complementarity with other key stakeholders at the local, regional, national and/or international level).
- **By thematic areas.**

If evaluation questions are clustered by transversal and/or thematic areas – the recommended option in the case of the Directorate-General for Neighbourhood and Enlargement Negotiations (DG NEAR – one or more evaluation criteria would be covered at the same time within each area.

GUIDANCE FOR DRAFTING INDICATIVE EVALUATION QUESTIONS

When formulating the indicative evaluation questions to be included in the terms of reference (ToR) of the evaluation, evaluation managers should follow this guidance.

- Ensure **consistency** between the evaluation questions and the evaluation objectives and scope as described in the evaluation ToR.
- Avoid excessively generic formulations; **tailor** the evaluation questions to the needs of the evaluation, the specificities of the intervention to be evaluated and its context.
- Use straightforward, **plain language**.
- Opt for **open-ended** rather than closed-ended questions, such as the following:
  - **When organised by evaluation criteria:** Which factors critically influenced the efficient implementation/delivery of support?
  - **When organised by transversal cluster:** To what extent has the intervention been designed and implemented so as to maximise EU (i.e. European Commission plus European External Action Service plus EU Member State) cooperation potential and EU added value?
  - **When organised by thematic clusters:** To what extent has the intervention contributed to improvements in sustainable production practices?

TABLE 3.1.2 Examples of evaluation questions by evaluation criteria

Criterion	Example
Relevance	How does the intervention presently respond to the needs of the ministry for transport?
Coherence	How coherent is the intervention with other EU actions in the country?
Efficiency	To what extent have the outputs been produced/delivered in a cost-efficient manner?
Impact	To what extent has the intervention contributed towards reinforcing regional integration?
Sustainability	To what extent has the intervention helped generate effect X in such a way that it lasts after the end of the intervention? To what extent has the partner government acquired the skills to continue implementing the services provided by the intervention once it comes to an end?
EU added value	To what extent does an intervention in the tourism sector add value to what Member States are doing?

- Construct **clear hypotheses** to be tested by the evaluation (e.g. how and to what extent does the provision of electricity in community X affect gender equality?).
- If possible, relate the question to **available evidence** (e.g. how and to what extent does the X percentage increase in children's participation in sports activities as referenced in document Y contribute to improved achievement in school?). If possible, address a **known gap** (e.g. as a follow-up to the previous question, add 'when answering the question, the team will assess whether and how gender/minority/income differences affect participation and school performance').

SEE: *The step-by-step guidance in the How-to Guide on evaluation questions in the Evaluation wiki for detailed information.*

NOTE: *The evaluation questions must be agreed upon with the reference group. Ideally, this should happen before finalising the ToR but, if not possible, this should happen during the inception phase.*

NARROWING DOWN THE NUMBER OF QUESTIONS

In choosing evaluation questions (see Figure 3.1.1), **less is more**. As few questions as possible should be selected to keep the focus of the evaluation clear and sharp. Overburdening an evaluation with questions and criteria generally results in poor evaluation quality, as evaluators will have less time to address each properly. Additionally, as a practical consideration, the wider the focus of the analysis, the higher the evaluation cost will be.

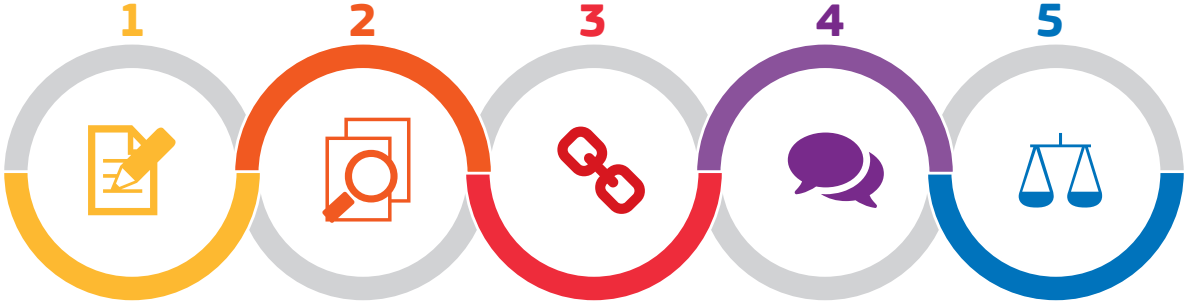
A maximum of 10 (12 for strategic and budget support evaluations) evaluation questions should be agreed upon, but a good evaluation can have as few as 5 or 6 well-tailored questions.

NOTE: *It is useful to number the evaluation questions; this will simplify referencing them during their finalisation and reporting.*

Three key factors can help in determining the utility and relevance of an evaluation question.

- **The question is raised by a relevant evaluation stakeholder.** Such stakeholders include members of the public, European Commission (EC) service staff (particularly those participating in the reference group) or a key informant consulted by the evaluation manager.
- **The answer is useful to know.** A question is particularly useful if:
  - the evaluand or one of its aspects is innovative, and several actors expect validation;

FIGURE 3.1.1 How to choose the evaluation questions



1. Establish a preliminary list of questions based on intervention logic/ theory of change, evaluation criteria, initial needs and key issues
2. Assess how and by whom the answer will be used
   - Check that the answer will be delivered on time to be useful
   - Check that the answer is not known already
3. Check that the answer is likely to be sufficiently robust
4. Discuss the choice of questions with the reference group so they can be agreed upon ideally before finalising the ToR
5. Choose and consolidate questions as impartially as possible

- the evaluation's findings will be ready in time to help make a planned decision;
- the evaluation's findings will be ready in time to feed into a planned public debate.

- **The answer is not known.** It is a waste of resources to include a question for which another evaluation, audit or study has already provided an answer.

Take the following criteria into consideration when attempting to narrow down the number of evaluation questions (DG NEAR, 2016):

- **There is a genuine interest in knowing the answer and using the resulting knowledge.** An evaluation should avoid a compliance attitude – carrying out an evaluation because it is mandatory to do so or simply because it was planned – or pressuring the evaluation team to provide desired answers to the evaluation questions.
- **Feasibility.** Can the answer to the question be found in a reasonable amount of time and within the limits of available resources?
- **Resources.** Are the time, financial and human resources assigned to the evaluation and its management appropriate to the task and are sufficient data available?
- **Openness to criticism.** Are the commissioning entity and other key stakeholders willing to accept unexpected or unfavourable answers – especially when evaluation reports are made public?
- **Ownership.** This refers to the level of engagement of key stakeholders in the formulation of the evaluation questions.
- **Consensus.** Is there a strong consensus around the need for the evaluation question? Evaluation managers and evaluators give higher priority to questions that are relevant to the greater number of stakeholders.

JUDGEMENT CRITERIA

Judgement criteria are derived for each evaluation question and specify aspects of the evaluand that will allow its merits or success to be objectively assessed. They inform on how to judge, not on what is judged and guide the evaluation team on **how to answer each evaluation question** after having collected and analysed all relevant data. Thus, each judgement criterion should be accompanied by a target level and one or more indicators.

Judgement criteria help to:

- avoid subjectivity and to formulate judgements on accepted terms;
- improve the transparency of the evaluation by making the judgement explicit;
- structure the answers to the questions asked, since the judgement criteria will determine the indicators and, more generally, the nature of the data collected (see Section 3.3) and the type of analysis (see Section 3.4).

All the evaluation questions relate to one or more judgement criteria. The following is an example of an evaluation question:

> To what extent has EC support improved the capacity of the primary educational system to enrol pupils from underprivileged groups without discrimination?

Like most evaluative questions, it has two parts.

- **What is being judged:** In this case, EC support.
- **The way of judging:** Has EC support, for example, improved the capacity of the primary educational system to enrol pupils from underprivileged groups without discrimination?

The judgement criteria develop and specify the second part of the question, for example:

- capacity of the primary school system to enrol pupils from ethnic minority X satisfactorily;
- capacity of the primary school system to enrol pupils from disadvantaged urban areas satisfactorily.

The judgement criteria derive from the question; the following illustrates this for the first judgement criterion cited above (capacity of the primary school system to enrol pupils from ethnic minority X satisfactorily).

- It concerns the way of judging and not what is judged. This is why the beginning of the question concerning EC aid has been removed.
- It specifies the type of success to be evaluated, that is, an improvement in the capacity of the primary school system to enrol pupils from underprivileged groups without discrimination, and specifically pupils from ethnic minority X.
- It emphasises the judgement and not the causality analysis. That is why the terms 'to what extent… has it improved' have been removed.
- To be used in practice, each judgement criterion has to be accompanied by one or more indicators.

SEE: *The discussion on Judgement references on Capacity4dev's Evaluation methodological approach wiki for detailed information.*

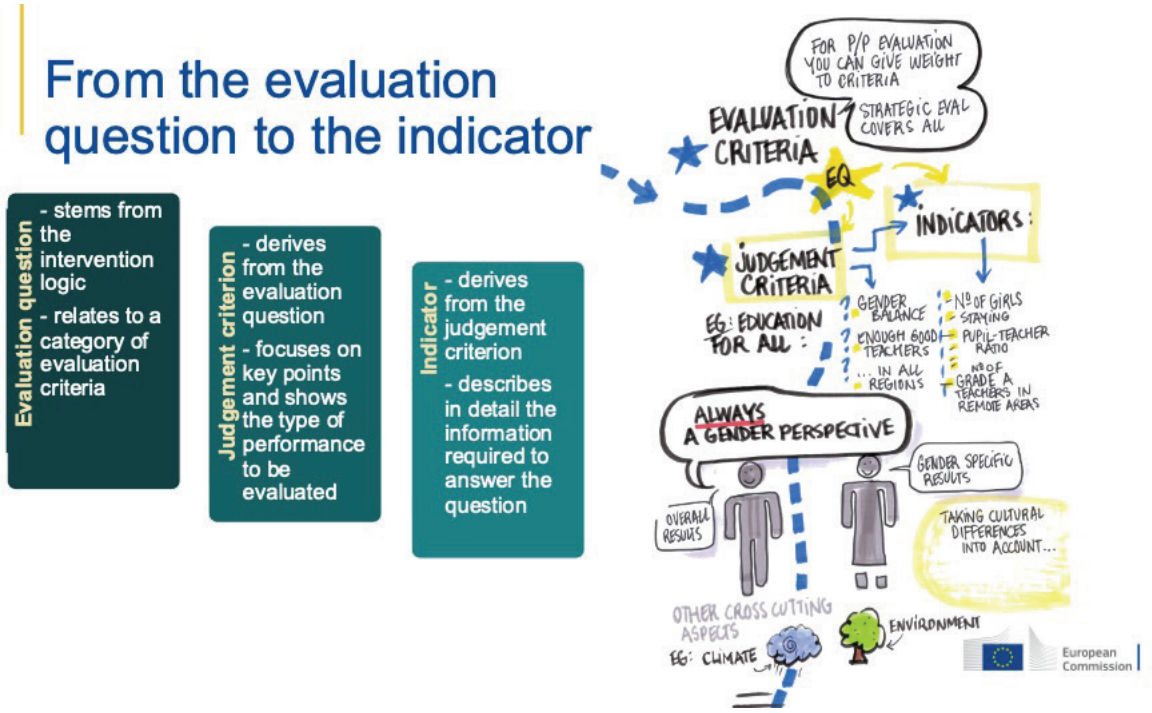
Key points to keep in mind when developing judgement criteria follow.

- To avoid availability bias and a reliance on existing information, always **define the judgement criterion** before selecting either an existing indicator or creating a new indicator.
- **Discuss judgement criteria with the reference group** so a diversity of viewpoints relevant to the intervention can be taken into account.
- To optimise data collection, **define a limited number** of judgement criteria for each question. This also takes into account users' capacity to absorb information.
- Where relevant, **explain any gaps** between the criteria used to formulate the judgement at the end of the evaluation process and those identified in the desk activities during the evaluation's interim phase.

INDICATORS

As part of an evaluation, it is often important to either develop or use existing indicators (see Figure 3.1.2) or measures of implementation and/or results. Indicators qualify the judgement criteria and help to specify the type of information to be collected/analysed. They provide accurate and non-ambiguous information that can be understood in the same way by all evaluators and users. The Better Regulation Guidelines (EC, 2021a) specify that indicators should be RACER (relevant, accepted, credible, easy, robust)

FIGURE 3.1.2 Moving from the evaluation question to the indicator



and may be quantitative or qualitative. They may be based on the logframe indicators, but not exclusively.

Terms that are commonly associated with measurements include:

- **baseline value**, which is the value of an indicator at the outset of an intervention (e.g. number of people living below the poverty line at the start of the intervention);
- **target**, which is the value of an indicator expected to be achieved at a specified point in time, generally by the end of the intervention (e.g. number of people living below the poverty line by the end of the intervention);
- **index**, a set of related indicators which intend to provide a means for meaningful and systematic comparisons of performance across interventions that are similar in content and/or have the same goals and objectives (e.g. the Human Development Index);
- **standard**, which is a set of related indicators, benchmarks or indices which provide socially meaningful information regarding performance (e.g., the international poverty line).

SECTION 3.2

Evaluation design

This section provides an overview of different evaluation approaches, methods and methodologies:

- **approach** is understood to be the underlying logic by which an evaluation addresses attribution and causation;
- **methods** refer to the tools or techniques that can be used in support of an evaluation approach;
- **methodology** indicates a system by which evaluation methods are organised.

This section looks at the key factors that need to be taken into consideration when selecting the methodology to guide an evaluation such as the nature of what is being evaluated, the type of evaluation, and the resources and constraints. The different types of evaluation questions – causal, descriptive and normative – and their implications for evaluation design are then discussed. The remainder of the section then describes a wide variety of theory-based, participatory and other approaches to evaluation design.

3.2.1 Factors to consider in making design decisions

A good evaluation design should consider three sets of factors: (i) the nature of what is being evaluated, (ii) the type of evaluation, and (iii) the resources and constraints.

NATURE OF WHAT IS BEING EVALUATED

Different types of evaluands need different types of evaluation designs. The evaluation design, while considering the purpose and intended uses of the evaluation, should consider what is currently known about the evaluand in terms of its level of predictability, change trajectory, variability, other contributing factors and likely follow-up interventions.

NOTE: *Gender should also be taken into account. See Box 3.2.1.*

Level of predictability

Evaluand outcomes and impacts that are well understood and have a relatively lean and simple intervention logic can be reasonably predicted to a certain extent if implementation is adequate. For example, building a local dispensary or hospital has positive effects on the health of local communities.

Predictability becomes more challenging for complex evaluands, such as those in fragile or rapidly changing contexts or new and innovative interventions, where the intervention logic changes in response to changed circumstances or changing understanding about what is needed or what works in a particular situation, and where adaptive management has changing information needs. Changes in circumstances frequently occur that will affect the choice and timing of data collection.

Change trajectory

This is the graph of expected results compared to a baseline. Some interventions have a long lag time before any changes in impacts are visible; others have a slow start and then exponential growth; while some show initial results which fade quickly. These all have implications for when data collection should best be undertaken, or how data should be interpreted if it is not possible to get longitudinal data (data collected over time). For example, an evaluation may discover that people have high levels of knowledge soon after attending a training session, but if this training is not reinforced, the knowledge will most likely fade over time. Measuring knowledge levels soon after a training course would therefore not be a good predictor of the level of sustained knowledge.

BOX 3.2.1 Ensuring gender-responsive evaluation

To ensure that evaluations are gender sensitive, the following checklist can be applied. Ideally, all questions should be answered positively.

- Is the evaluation process participatory? Does it provide for an equitable participation of women/girls and men/boys and of their respective organisations, including those that are marginalised or disadvantaged, and persons with disabilities?
- Have gender-sensitive indicators been developed to measure both qualitative and quantitative results, at all levels of the results chain?
- Will gender-sensitive indicators be used in this evaluation?
- Was an initial assessment done to define which mix of data collection and analysis methods will be used to address data gaps and weaknesses with respect to gender equality?
- Do sex- and age-disaggregated baseline data exist?
- Are data-gathering and analysis tools (both qualitative and quantitative) designed to disaggregate and measure the results of the intervention for both women/girls and men/boys?
- Is gender (and age, ethnicity, sexual orientation, disability etc., if relevant) included among the criteria used to build up the consultation sample?
- Is the methodological approach flexible enough and is sufficient time allowed to respond to constraints and challenges of the informants, taking into consideration their gender and age?
- Does the methodology take into account the ethical and safety measures necessary to protect the informants?

Variability

The more differences or variations that exist in the context, implementation modalities or target populations of an intervention, the more varied the range of methods used in an evaluation design will need to be. With increasing attention to equity issues and the principle of 'no one left behind', it is important to ensure that evaluations adequately represent the relevant range of stakeholder experiences. Strategic

evaluations face particular challenges in this regard due to the diversity of interventions that are included in a single evaluation.

Other contributing factors

The results of interventions are affected by many other factors, including interventions that were implemented by other organisations as well as external factors (political unrest, economic crises, climate change–related weather events etc.). The evaluation design needs to adequately address these to better understand what works in what contexts, and to support learning that can be translated to new contexts.

For example, the level of political unrest can affect participation in community activities – so an evaluation might seek to understand whether sites with low levels of participation/progress have been mostly in areas affected by political unrest, or whether there are some areas that have managed to achieve high levels of participation despite these unfavourable circumstances, as these might offer lessons for other locations.

TYPE OF EVALUATION

The evaluation design should also take into account the type of evaluation in terms of its intended uses, the types of questions being asked, and its level of complexity such as multiple components or levels in implementation.

Intended uses of the evaluation

The evaluation design is intended to answer a number of evaluation questions, which have been developed to respond to the needs of the primary intended users. For example, if the primary intended use of the evaluation is to provide accountability to European Union (EU) institutions and citizens, then it will focus particularly on analysing implementation and achievement of objectives. If, on the other hand, the evaluation is meant to inform future interventions, the evaluation questions will focus on drawing lessons from experience that can be used in the definition of new interventions in the same field.

Types of questions being asked

The design should take into consideration the formulation of the evaluation questions, particularly in terms of (i) what would be considered credible evidence in answering them and (ii) when the answers are needed.

- The evaluation design should identify relevant and credible sources of evidence that will allow the evaluation questions to be answered in a meaningful way.
- In many cases, evaluations have a relatively short time frame from initiation to when findings are needed. This precludes some potentially useful data collection methods, in particular longitudinal research which involves gathering data over a longer time frame – for example, by tracking participants' experiences in an intervention and then assessing which outcomes and impacts are sustained years afterwards.

RESOURCES AND CONSTRAINTS

Evaluations need to be designed to fit with available resources and constraints. These include the availability, relevance and quality of existing data; the availability of funding; and the time requirements on operational staff and stakeholders such as intended beneficiaries or partner organisations. Constraints would include security risks in conducting primary data collection in areas that are conflict-affected.

> **NOTE:** ***Primary data*** *refers to the first-hand data gathered by the evaluation team;* ***secondary data*** *means data collected previously by someone else.*

Strategic evaluations face particular challenges, as there is less opportunity to do primary data collection across all the concerned areas of focus and sites/locations to fill gaps in existing secondary data.

The example in Box 3.2.2 illustrates how focused primary data collection adds value to the analysis of secondary data. This allows for a more complete understanding of what has happened and why.

BOX 3.2.2 The value added of collecting primary and secondary data

The 2020 External Evaluation of the European Union's Cooperation with Myanmar (2012-2017) covered a number of different sectors, including education. One of the intended outcomes in education was to modernise approaches to teaching and learning from rote learning and memorisation to active learning. The evaluation was able to draw on existing evidence from monitoring systems and previous evaluations and research studies about many of the steps in the results chain – the number of teacher training activities that had been completed, the content of these activities, the attitudes of participants to what they had learned, changes in teacher behaviours and improvements in student learning. These data included official reporting on activities and test scores, surveys of participants and direct observation of a large sample of teaching which showed that more than 35 per cent of teachers had adopted the new practices. The evaluation therefore focused primary data collection on key informant interviews to confirm and explain these findings. In particular, the evaluation sought to explain the barriers to teachers adopting and sustaining the new practices, including systemic disincentives such as the continued focus on memorisation and recall in examinations and scholarship competitions, and resultant lack of parental support for the new teaching methods.

3.2.2 Design by type of evaluation question

As noted above, a core factor influencing the evaluation design will be the agreed-upon evaluation questions. The different types of evaluation questions and their implications for evaluation design are discussed in this subsection.

Table 3.2.1 presents various types of evaluation questions along with examples and the associated evaluation criteria. Note that answering each question will require different methods and tools.

DESCRIPTIVE QUESTIONS

Descriptive questions ask **what has happened** and require evaluators to define, observe and measure change – often from the point of view of various stakeholders. These questions pertain to positive and negative changes, be they expected or unexpected, directly or indirectly linked to the intervention. Typical examples follow.

- What was the situation before the start of the intervention (baseline)?
- What is the situation now?
- What happened between now and then?
- How do changes differ for each area/sector/affected group?

These questions need answers that contain the definitional information about the evaluand or describe some particular events. In such cases, using as many sources of data as possible in the evaluation would help in finding the most appropriate answers (i.e. data triangulation – see Box 3.2.3).

NORMATIVE QUESTIONS

A normative (or valuing) question is one that asks **what should be** rather than one that is designed to determine an objective outcome or condition, such as 'how much' or 'yes' or 'no'. The purpose of a normative question is to define what is best in a given situation. For example, a question that asks what the unemployment rate in country X should be is a normative question, or whether the intervention was worth the cost.

Evaluation designs need to answer normative questions in a way that is systematic, transparent, and defensible. If the norms and values are not explicitly addressed in the evaluation design, there is a risk that arbitrary judgements will be made about whether the intervention is satisfactory on the basis of simple comparisons between before and after implementation scenarios, or by comparing a particular site to a national average.

There is a four-part logic involved in answering a normative/valuing question (though these could also apply to other types of questions):

TABLE 3.2.1 Types and examples of evaluation questions

Type	Description	Example	Evaluation criteria
Descriptive	• Ask what has happened and require evaluators to define, observe and measure change, often from the point of view of various stakeholders. • Questions pertain to positive and negative changes, be they expected or unexpected, directly or indirectly linked to the intervention.	• What was the situation before the beginning of the intervention (baseline)? • What is the situation now? • What happened between now and then? • How do changes differ for each area/sector/affected group?	Effectiveness, particularly those about changes brought about by the project
Normative	• Ask how an intervention fares against a criterion. • When selecting criteria, evaluation managers always have to explicitly state the rationale for their choice in the documents accompanying the start of an evaluation.	Is the intervention worth the cost?	Relevance; coherence
Causal	• Shed light on whether an intervention works and on how it works, for whom and under what circumstances. • Produce knowledge that can be used to improve interventions, to identify indicators, to understand problems and fix them, and to launch new, effective initiatives in the future.	• Has the intervention produced these changes (or stopped change?) • What contributed to the changes?	Impact; effectiveness

SOURCE: DG NEAR (2016).

- the criteria to be used (domains of performance);
- standards that apply (levels of performance);
- evidence required (relevant and credible data about performance);
- synthesis (combining data to produce an overall judgement).

Two broad approaches to establishing what is 'good' performance in terms of implementation processes or results are the use of established criteria and standards, targets or benchmarks, and the development of rubrics.

- **Criteria, standards, targets or benchmarks.** Sometimes there are already established criteria

BOX 3.2.3 Data triangulation

Triangulation is about using a complementary combination of traceable data sources. It is an important part of good design in answering evaluation questions. The appropriate choice of data collection methods and their effective implementation is also important. For example, in cases where it is difficult to speak freely about some (sensitive) issues in a group interview situation or where privacy for an interviewee cannot be achieved, further methods to gather evidence could be used to validate findings through different sources. For example, the 2015 External Evaluation of Sustainable Rural Development in the Refugee-Affected and Hosting Areas of Pakistan Programme demonstrates the importance of triangulating data from different sources when assessing the credibility of evidence. The number of feasibility studies of proposed projects was reported to be the same as the number of implemented projects; therefore, the assumption was made that all projects subject to a feasibility study had been implemented. However, these official data were contradicted by interviews with a provincial team where social organisers stated that, on average, 10 to 20 per cent of the projects were refused following a feasibility study.

for evaluation and levels of performance by which to judge the quality of the intervention. These might relate to policy objectives, international agreements or other official sources. Specific targets will often have been established for the intervention when its monitoring framework was set up.

- **Rubrics.** A rubric is a framework that sets out criteria and standards for different levels of performance and describes what performance would look like at each level (see Box 3.2.4). The term was originally used to refer to how student work would be graded, but it has since been expanded and used for evaluating interventions. Another label for it is a global assessment scale.
- **Most significant change.** Most significant change (discussed at length in Subsection 3.2.6) is a form of participatory monitoring and evaluation that involves the collection and selection of stories of significant changes that have occurred in the field. It answers questions about what is valued by different stakeholders and generates data for developing and testing intervention logic.

An evaluation design is likely to include a different design for different types of questions, as the example in Box 3.2.5 shows.

CAUSAL QUESTIONS

Causal questions ask about **why results occurred**, assessing the connection between an intervention and outcomes and impacts (intended or unintended). Different types of causal questions might be appropriate:

- Did the evaluand make a difference?
- For whom, in what situations, and in what ways did the intervention make a difference?
- How much of a difference did the intervention make?
- To what extent can a specific impact be attributed to the intervention?
- How did the intervention make a difference?

Causal questions usually recognise that multiple factors contribute to producing the changes that have been observed – a combination of the intervention that is being evaluated and the context in which it has been implemented, including other interventions that either help or hinder it. Causal questions aim to allow causal claims or evidence claims to be made. **They seek to assess the extent to which an observed change can be attributed to a given intervention.**

BOX 3.2.4 Example of using a rubric to answer a normative question – Midterm review of Promotion of Inclusive and Sustainable Growth in the Agricultural Sector: Fisheries and Livestock in Cambodia (2015)

To answer a question about the outcomes of capacity-building activities, the evaluation team used the following rubric to assess staff capacities across a number of component tasks at the start and end of the intervention. Scores were based on a synthesis of data from different sources: self-assessment; assessment by long-term technical assistance, through six-monthly and technical reports and interviews; and assessment by short-term technical assistance. Changes in scores provided part of the answer to an evaluation question about the extent to which capacity of relevant departments had been developed.

5	Well managed autonomously
4	Well managed with punctual support
3	Well managed with regular support / partially managed correctly with punctual support
2	Partially managed correctly with regular support
1	Weakly managed with extensive support
0	Totally delegated to external technical assistance / unable to manage

BOX 3.2.5 Using different designs to answer different normative questions

The Management of Protected Areas to Support Sustainable Economies project (2017) evaluation used three different ways to answer normative/valuing questions.

- **Valuing against established standards.** A structural engineers' report on eco-tourism facilities developed by the project assessed the physical infrastructure against established engineering standards.
- **Comparison with benchmarks.** The overall project objective was to fulfil international environmental agreements and, in so doing, support sustainable development. Results were assessed against the guiding principles of the Turks and Caicos Islands Environmental Charter, the Convention on Biological Diversity Aichi targets and aspects of the Specially Protected Areas and Wildlife Protocol and the mandate of the National Trust.
- **Using a rubric to assess quality of results.** Qualitative and quantitative data were synthesised to generate a rating (very good, good, problems, serious deficiencies) of the project results in developing sustainable eco-tourism facilities. For example, in the Turks and Caicos territory the results were rated as very good, it was noted that:

> 3 of 4 sites operationalised and monetised. All sites well maintained, show continuous visitor growth and create sustainable, unrestricted income for TCNT [Turks and Caicos National Trust]. While 1 visitor facility could not be realised and was removed from the scope (mainly due to influences outside the project's control), the results for this intervention logic can still be evaluated as 'very good' since they are functional and create the intended impact.

The Organisation for Economic Co-operation and Development (OECD) defines attribution as the 'ascription of a causal link between observed (or expected to be observed) changes and a specific intervention'. Measurement of attribution relies on establishing causality – that is, evidence that an intervention directly caused the observed outcomes; it is usually assessed by means of a counterfactual – what would have happened if the intervention had never occurred.

This definition does not require that changes are produced solely or wholly by the intervention or policy under investigation. Rather, it takes into consideration that other causes may also have been involved – for example, other interventions/policies in the area of interest or certain contextual factors (often referred to as 'external factors').

The key challenge in impact evaluation is that the counterfactual cannot be directly observed and must be approximated with reference to a comparison group. There are a range of accepted approaches to determining an appropriate comparison group for counterfactual analysis, using either prospective (ex ante) or retrospective (ex post) evaluation design. **Prospective evaluations** begin during the design phase of the intervention, involving collection of baseline and endline data from intervention beneficiaries (the treatment group) and non-beneficiaries (the comparison group); they may involve selection of individuals or communities into treatment and comparison groups. **Retrospective evaluations** are usually conducted after the implementation phase and may exploit existing survey data, although the best evaluations will collect data as close to baseline as possible to ensure comparability of intervention and comparison groups.

There are three ways to develop a counterfactual:

- Experimental designs construct a comparison (control) group through random assignment.
- Quasi-experimental designs construct a comparison group through matching, regression discontinuity, propensity scores or other means;
- Non-experimental designs look systematically at whether the evidence is consistent with what would be expected if the intervention was producing the observed impacts, and/or whether other factors could provide an alternative explanation.

Experimental designs

Experimental designs create a control group – a group of **randomly assigned participants** who do not receive the intervention – and compare the outcomes of this control group with those of the treatment group – that is, the group who are recipients of the intervention. This kind of design must be established before the intervention begins so that people or sites can be randomly assigned and changes tracked. It is most suitable in situations where a discrete, standardised intervention is being evaluated, and where the evaluation is sufficiently large, and variability sufficiently small, that variations in effects can be adequately covered so the evaluation can produce valid and useful findings.

Not all interventions are amenable to being evaluated using experimental methods. Experimental methods are difficult to use in evaluating development interventions, as they often involve complex social and economic processes that cannot be easily manipulated or controlled.

Experimental methods should be used in evaluations when it is possible to randomly assign the intervention to households or individuals. The evaluation should also be designed in such a way that, as much as possible, the intervention is the only difference between the treatment and control groups. Otherwise, it is not possible to attribute any differences in outcomes to the intervention.

An example of an intervention that is suited to experimental evaluation methods is one that provides cash transfers to households in developing countries. This type of intervention can be evaluated using a randomised control trial (see Box 3.2.6), in which some households are randomly selected to receive the cash transfer and others are not. An intervention that provides infrastructure development, on the other hand, cannot be evaluated using a randomised control trial because it is not possible to randomly assign infrastructure projects to households or communities.

Quasi-experimental designs

In experimental design (true experiments), an evaluator has full control over the events of interest. In order to evaluate whether an intervention works or not, the evaluator can design a 'laboratory' where 'experiments' are performed, and all the relevant factors are controlled. In practice, this option is rarely applied in evaluation (it is more common in scientific research) as it requires significant resources. The complete isolation of the studied case may also be impossible. Hence, **quasi-experimental designs** are more common in the evaluation arena. In this case, the evaluator has less control over the factors and experimental set-up. This means that there is no randomisation. It also requires strong statistics and computational skills, as the differences between populations are normally subject to complex calculations. Typically, an evaluator develops a number of rival hypotheses and examines the differences between the outcomes in each test.

Quasi-experimental designs construct a comparison group as a counterfactual without randomisation. Ways of creating a comparison include matching participants (individuals, organisations or communities) with a non-participant on variables that are thought to be relevant. If the comparison group is not really comparable (and it can be difficult to be comparable in terms of all the key factors), then it will not provide a valid counterfactual. In some cases, it is possible to develop a hypothetical counterfactual – for example, using the baseline as an estimate of the counterfactual under conditions of no change, or asking key informants to estimate the counterfactual.

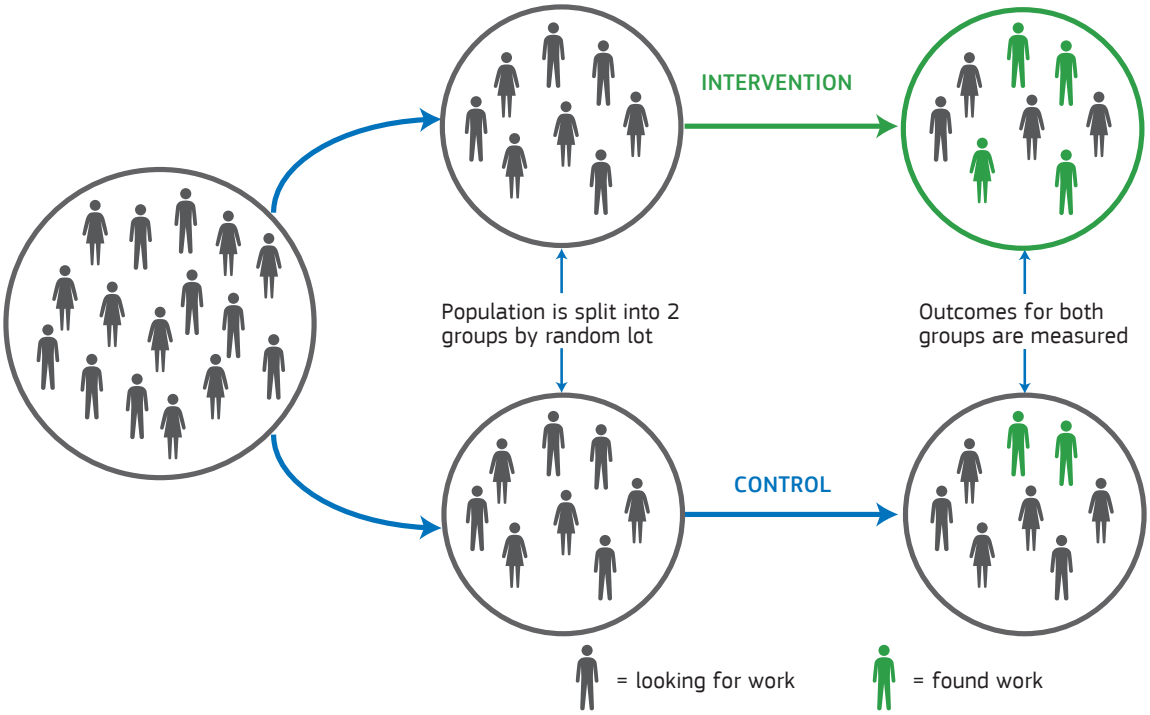
Non-experimental designs

Although evaluators can often measure whether an intended outcome has occurred or not, it is difficult to determine what outcomes are attributable to a specific intervention. Unless the target population is extremely narrow, it will be difficult to show attribution in evaluation. In addition, the complex and comprehensive nature of most interventions makes inferring causation extremely difficult because there may be multiple initiatives designed to support each other with multiple activities which may not have explicit, measurable objectives. In addition, interventions operate in complex social environments; in most cases, there are many other factors at play in addition to those resulting from a given intervention's activities.

BOX 3.2.6 Randomised control trials

Randomised control trials are a type of experiment that aim to reduce the influence of all factors other than those of the intervention being evaluated (see Figure B3.2.6.1). This is done by randomly assigning cases into treatment and control groups. The treatment group receives the intervention, the control group does not. The outcomes of the two groups are then compared. **Blinded** trials are experiments where the participants do not know they are receiving the treatment. **Double-blinded trials** are where neither the participant not the experimenter (evaluator) knows who is receiving the treatment. The aim is to reduce the biasing effects of expectations on outcomes. Blinding is common in drug trials but difficult, if not impossible, in most development contexts where the nature of the intervention cannot be hidden.

FIGURE B3.2.6.1 Basic design of a randomised control trial testing a new back-to-work programme



SOURCE: Haynes, Goldacre and Torgerson (2012).

There are varied and dynamic variables affecting the environment within which the intervention or multiple interventions operate, such as socioeconomic, environmental, political and cultural factors. These usually cannot be isolated, manipulated or measured, making it extremely difficult to show attribution. **Change is seldom attributable to a simple factor.**

A scientific or quasi-scientific approach in which tests of statistical differences in outcomes between treatment groups and comparison (or control) groups may not always be possible or desirable. White and Phillips (2012) make the distinction between large *n* quantitative studies and qualitative middle/small *n* studies. The former use quantitative methods to examine statistical associations between the intervention and outcomes as probabilistic statements which can be interpreted causally when selection bias has been taken into account. The latter use qualitative approaches to examine necessary and sufficient conditions for an outcome's achievement. They go on

to categorise these non-experimental approaches as either:

- **theory-based approaches** (discussed in Subsection 3.2.3 and Subsection 3.2.4) that make causal claims based on plausible association and absence of other explanatory factors;
- **participatory approaches** (discussed in Subsection 3.2.5 and Subsection 3.2.6) that rely on stakeholders' assessment of intervention impact.

White and Phillips note that, although the latter do yield useful information, they should not be the basis for claiming impact.

3.2.3 Theory-based approaches: understanding the intervention logic

These approaches explicitly set out to discover the causes of observed effects with the goal of establishing beyond a reasonable doubt how an outcome or set of outcomes occurred. Most of these approaches emphasise the need to draw on the implicit theory of change or intervention logic underpinning an intervention, and to map out steps by which an evaluator can assess whether the theoretically proposed changes occurred as expected; whether the causes and assumptions set out in the theory of change or intervention logic varied; or whether the observed outcomes were a result, in part or whole, of other external factors. Before looking at some of the more commonly used theory-based approaches, the following provides an overview of the theory of change or intervention logic as it is referred to by the Directorate-General for International Partnerships (DG INTPA) and the Service for Foreign Policy Instruments (FPI).

> **NOTE:** *The intervention logic for budget support operations is included in the Annex to this handbook.*

From the above, it is clear that a reasoned intervention logic / theory of change is fundamental to the evaluation process. Although it is defined during the design phase of an intervention, the intervention logic will evolve throughout implementation. In many cases, the quality of the original intervention logic may be weak, or the context may have evolved to such an extent that the original intervention logic is now obsolete. In other cases, the intervention logic has not been updated to reflect changes made during implementation. In all these cases, the evaluation team will most probably need to 'reconstruct' the intervention logic to ensure that it adequately captures the planned change process – the hierarchy of expected results (outputs, outcomes, impact etc.) and the assumptions deemed necessary for the intervention to deliver as planned. In the absence of any intervention logic (unlikely, but possible), the evaluation team will need to draw on available documentation and initial interviews to reconstruct it from scratch.

The intervention logic can be described as a (narrative) description and/or a diagram summarising how an intervention is expected to deliver results. It describes the expected logic of the intervention or chain of events that should lead to the intended change. The intervention logic identifies the causal links between the outputs and the outcomes, and between the outcomes and the impact – also known as the **results chain** – as well as the **key assumptions** which underpin that change process.

The EU does not mandate a specific format or approach to use in designing/reconstructing the intervention logic. Two commonly used formats are described below: the **logical framework approach** and the **theory of change approach**.

THE LOGICAL FRAMEWORK APPROACH

The logical framework approach was developed in the late 1960s to assist the US Agency of International Development (USAID) with project planning. Most large international donor agencies, including the EU, now use some type of logical or results framework to guide project/intervention design and to inform evaluations.

Structure of the logical framework matrix

The standard European Commission (EC) logical framework (or logframe) consists of a matrix with

FIGURE 3.2.1 **EC model of the logical framework**

six columns and four rows which summarises the key elements of the intervention (see Figure 3.2.1):

- **The hierarchy of objectives.** The first column, also known as the results chain, captures the intervention's development pathway or planned change process. Each result should be explained by the result immediately below. Although different donors use different terminology, a logframe typically summarises the following in its first column:
  - the overall objective/impact;
  - the specific objective(s)/outcomes;
  - the outputs;
  - the activities (generally a summary of key clusters of activities rather than an exhaustive list of activities which are better captured in a work plan).

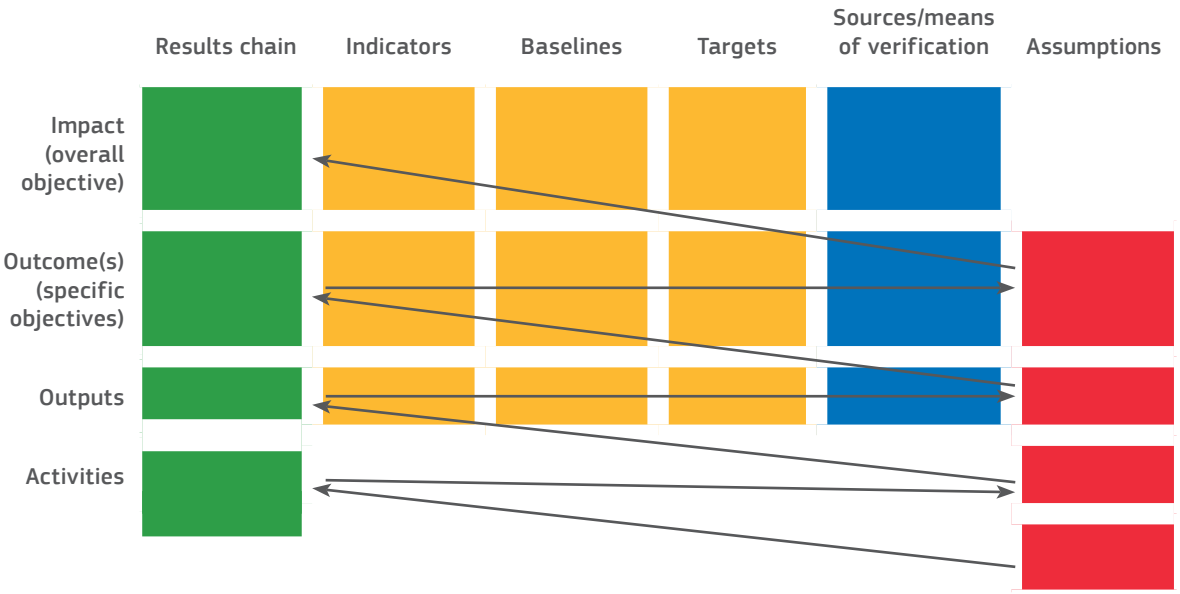
The second, third, fourth and fifth columns provide the basic information that will allow the corresponding results to be monitored, and consist of the following:

- **indicators** – a quantitative or qualitative measurement which provides a reliable way to measure changes connected to a given result;
- **baseline** – which records the value of the indicator before the intervention starts;
- **target** – the desired end point for each indicator, which is generally planned to be achieved in the last year of an intervention;
- **sources/means of verification** – which identify where the data for the corresponding indicators can be sourced.

The final column lists the **assumptions**. These are the external factors or conditions outside of the intervention's direct control that are necessary to ensure the intervention's success; these must hold for the results chain to materialise as planned. The assumptions should be formulated based on the context analysis and the risk assessment during the design phase and should be part of the monitoring system, as they will change over time. The assumptions can be environmental, contextual, causal and operational assumptions.

The articulation between the first column of the logframe (the results chain) and the last column (the assumptions) represents (a summary of) the intervention logic. This articulation is reflected in Figure 3.2.2. If the activities are carried out as planned and the corresponding assumptions hold, then the outputs will be delivered as planned. Similarly, if the outputs are delivered and the corresponding assumptions hold, then the outcomes will be delivered as planned; if these outcomes are delivered and the

FIGURE 3.2.2 EC logical framework showing articulation between results and assumptions



corresponding assumptions hold, then the impact will be delivered. This articulation is also referred to as the **vertical logic** of the logframe. The **horizontal logic** is the indicators, baselines, targets and data sources for each planned result.

The logframe thus provides a compact summary of the **intentions** of an intervention and how its **implementation and progress** towards planned results can be tracked and measured. The logframe thus not only serves as a key management tool but also as a key tool for evaluators, allowing them to assess the extent to which an intervention is delivering or has delivered as planned. In those cases where indicators are adequately tracked by management, the resulting information is of great value to evaluators.

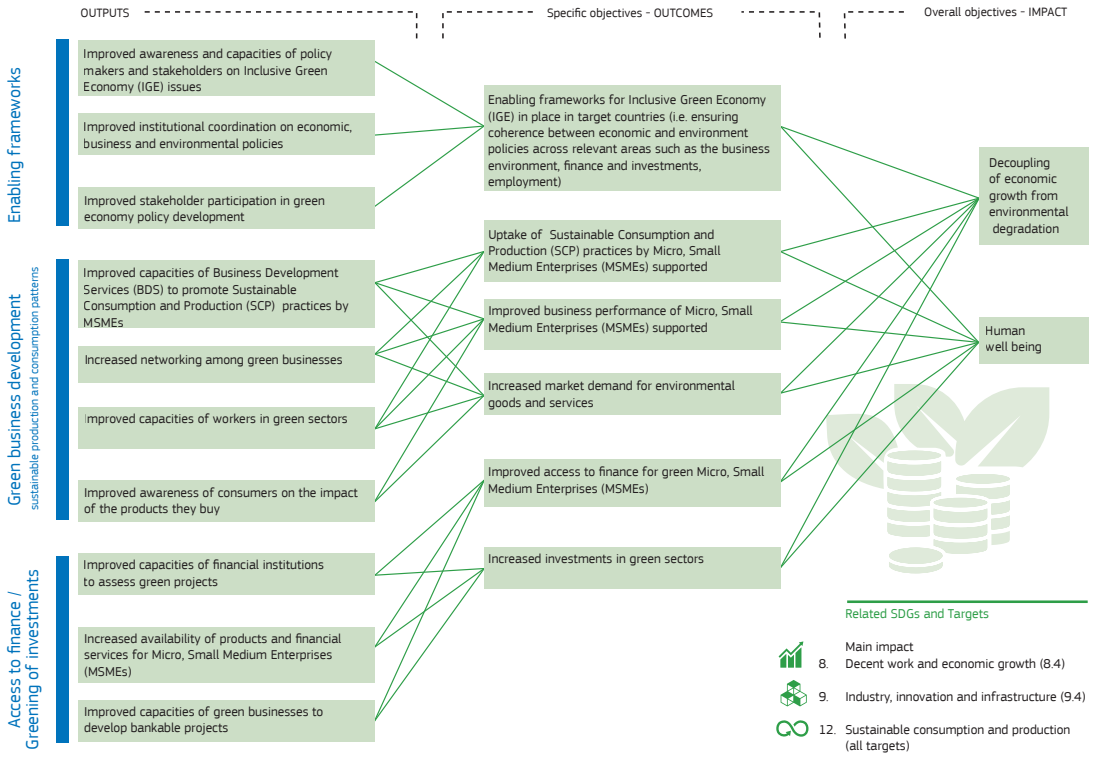
Challenges in using logframes

Despite its popularity and obvious strengths, there have been a number of criticisms levelled at the logframe approach which can be summarised as follows:

- **overly simplistic representation** of the change process, which in reality is far more complex than can be captured by a logframe;
- **insufficient detail as to which activities are contributing to which outputs**, and which outputs are contributing to which outcomes etc., particularly when an activity may be contributing towards more than one output, or an output may be contributing to more than one outcome – that is, the representation of change in the logframe is too linear;
- the need to **categorise all events as one of five types** (inputs, activities, outputs, outcomes, impact);
- the need to **condense all surrounding contextual influences into a single type** (assumptions) – the logframe is incapable of capturing the 'bigger picture', the external factors that may indirectly affect performance;
- the **unidirectional nature of the change process** – that is, the absence of any feedback processes;
- the difficulty of **using a logframe to communicate an intervention to those not familiar** with its structure.

Some of the shortcomings of logframes can be overcome by using complementary tools such as **diagrammatic representations of the intervention logic**. In these diagrams, labelled nodes typically represent events described in the left column of the logframe (the results chain), and links connecting the nodes typically represent the expected causal influence of one event on another – the vertical

FIGURE 3.2.3 Diagrammatic version of logframe clarifying expected causal connections between outcome and output



SOURCE: EC (2021b).

logic. The ways in which this can be done are many and varied. Figure 3.2.3 shows how the expected causal connections between outputs and outcomes can be made more explicit while retaining some of the structure of the logframe. Diagrams can also be much more complex, capturing a wide range of types of intermediate events and alternative causal pathways.

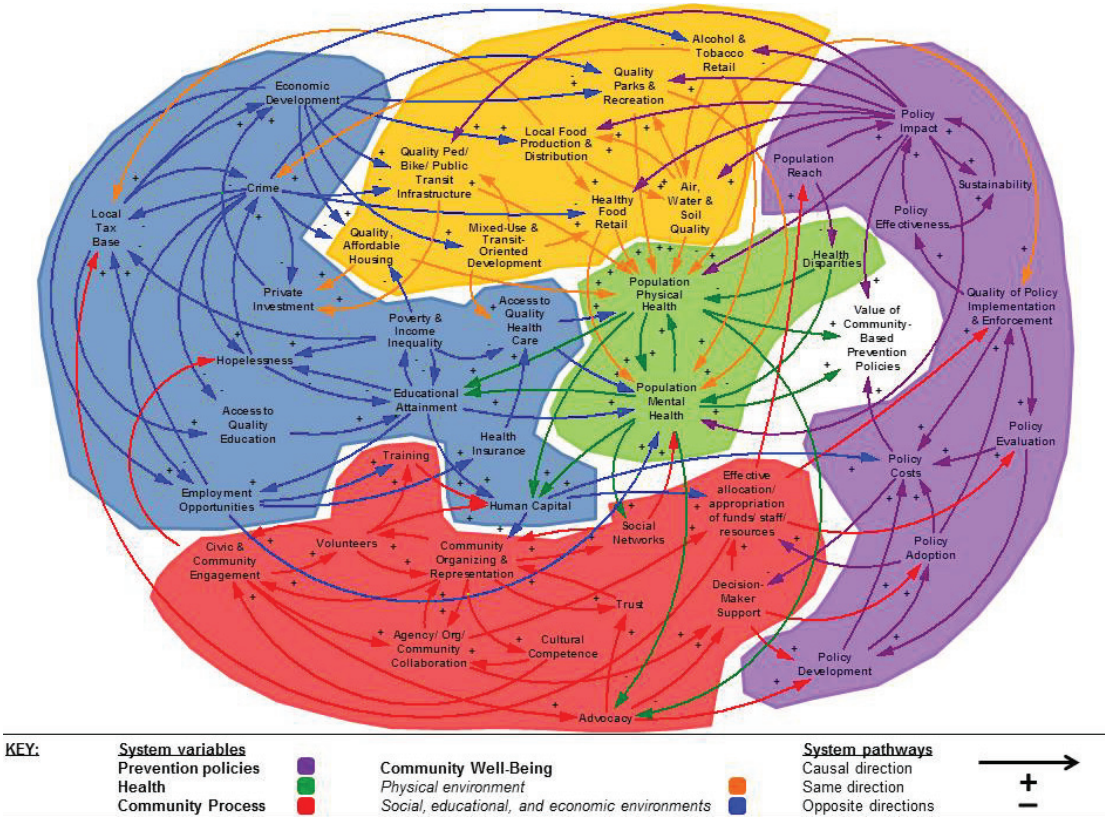
Diagrammatic representations do, however, have their **limitations**:

- They typically **leave out the details of the horizontal logic** – that is, how each event will be measured or observed and where that information will come from. Additional information needs to be provided in supporting text.
- The **description of the causal pathways can become very complex**, making it a challenge for an evaluation team to identify where to focus scarce resources and attention. The number of possible causal pathways can rise exponentially as new links are added into a diagram.
- While simple diagrams can make an intervention logic easy to communicate to non-experts, **complex diagrams can leave people confused**.
- Diagrams, even complex ones, **may not be able to capture the unpredictable nature of complex** strategies, policies, instruments, modalities or interventions in complex settings. Where events are densely connected by both positive and negative feedback links, it is not possible to identify by visual inspection alone what the net consequences will be (see Figure 3.2.4).

THE THEORY OF CHANGE APPROACH

Although the logframe approach was considered a significant advance, providing a framework through which the relationships between an intervention's components could be drawn out and articulated, the limitations described above led many evaluation experts to highlight the challenges posed when evaluating complex social or community change programmes – when it was not clear precisely what

FIGURE 3.2.4 Example of a causal loop diagram for value of community-based prevention policies



SOURCE: National Academy of Sciences (2012).

the programmes had set out to do or how, it made it difficult to evaluate whether or how they had achieved it (James, 2011). One organisation which began to focus on these issues was the US-based Aspen Institute and its Roundtable on Community Change. The work of the Roundtable led to the publication in 1995 of *New Approaches to Evaluating Comprehensive Community Initiatives*.

In that book, Carol Weiss, a member of the Roundtable's Steering Committee on Evaluation, hypothesised that a key reason complex programmes are so difficult to evaluate is that the assumptions that inspire them, and which form a crucial element of the logframe, are poorly articulated. She argued that stakeholders of complex community initiatives typically are unclear about how the change process will unfold and therefore pay little attention to the early and midterm changes that need to happen for a longer-term goal to be reached. The lack of clarity about the 'mini-steps' that must be taken to reach a long-term outcome not only makes the task of evaluating a complex initiative challenging but reduces the likelihood that all of the important factors related to the long-term goal will be addressed (Weiss, 1995).

Weiss popularised the term 'theory of change' to describe the set of assumptions that explain both the mini-steps that lead to the long-term goal, and the connections between programme activities and outcomes that occur at each step of the way. She challenged designers of complex community-based initiatives to be specific about the theories of change guiding their work and suggested that doing so would improve their overall evaluation plans and strengthen their ability to claim credit for outcomes that were predicted in their theory. She called for the use of an approach that seemed like common sense: lay out the sequence of outcomes that are expected to occur as the result of an intervention and plan an evaluation strategy around tracking whether these expected outcomes are actually produced.

Since the publication of Weiss's book, the use of planning and evaluation using theories of change has increased exponentially among philanthropies, government agencies, international non-governmental organisations (NGOs), the EU, the United Nations and many other major organisations in both developed and developing countries. This has led to new areas of work, such as linking the theory of change to **systems thinking** and complexity (which is discussed later in this chapter). Change processes are no longer seen as linear, but as having many feedback loops that need to be made explicit and understood.

Theory of change is essentially a comprehensive description and illustration of how and why a desired change is expected to happen in a particular context. It is focused on mapping out or filling in what has been described as the 'missing middle' between what a programme or intervention does (its activities) and how these lead to desired goals being achieved. It does this by first identifying the desired long-term goals and then works back from these to identify all the conditions that must be in place (and how these relate to one another causally) for the goals to occur. These are all mapped out in a framework that provides the basis for identifying what type of activity will lead to the results (outcomes) identified as preconditions for achieving the longer-term goal (impact). Through this approach, the precise link between activities and the achievement of the long-term goals is more fully understood. This leads to better planning, in that activities are linked to a detailed understanding of how change actually happens. It also leads to better evaluation, as it is possible to measure progress towards the achievement of longer-term goals that goes beyond the identification of outputs.

The advantages of the theory of change approach are mainly linked to its ability to capture various views and assumptions about the process of change. It helps in focusing on **how the change is triggered** – or not – by a particular strategy, policy, instrument, modality or intervention. A well-designed theory of change helps identify a wide range of underlying conditions on which achievement of the expected outcomes relies and lends itself to a **visual representation of how the change happens**; this can clarify the logic and facilitate communication among stakeholders.

As is the case for the logframe, an intervention's theory of change will be a core tool for the evaluation team. It serves as the blueprint for evaluation by identifying the framework against which the successes (or lack thereof) of a given intervention will be evaluated. Although it overcomes some of the criticisms made about the logframe, the theory of change also suffers from some of the weaknesses identified above for the diagrammatic versions of the intervention logic – the absence of detail as to how each event/change will be measured or observed; overcomplexity, making it challenging for an evaluation team to identify where to focus; and inability to capture the unpredictable nature of interventions in complex settings.

> **SEE:** *Connell and Kubisch (1998); Davies (2018); and Jackson (2013) for more on the theory of change approach.*

3.2.4 Commonly used theory-based approaches

Some of the most commonly used theory-based approaches to evaluation are:

- contribution analysis;
- global elimination methodology (GEM);
- realist evaluations;
- qualitative comparative analysis;
- comparative case studies;
- case analysis methods;
- process tracing.

> **NOTE:** *Examples of gender analysis tools and good practices that could be coupled with theory-based approaches include the Gender Results Effectiveness Scale developed by the United Nations Development Programme (UNDP); UNDP's Gender@work quadrants of change, highlighted in Good Practices in Gender-Responsive Evaluations (UN Women, 2020a); and UN Women's Rapid Assessment Tool to evaluate gender equality and women's empowerment results in humanitarian contexts (UN Women, 2020b).*

CONTRIBUTION ANALYSIS

The OECD defines contribution analysis as an approach for determining if – and how – an intervention contributed to an observed result, based on verifying the underlying theory of change.

Contribution analysis is an approach for assessing causal questions and inferring causality in real-life intervention evaluations. It offers a step-by-step approach that helps evaluators arrive at conclusions about the contribution the intervention they are evaluating has made (or is currently making) to particular outcomes and impacts. The essential value of contribution analysis is that it offers an approach designed to reduce uncertainty about the contribution the intervention is making to the observed results through an increased understanding of why the observed results have occurred (or not) and the roles played by the intervention and other internal and external factors. Contribution analysis is particularly useful in situations where the intervention is not experimental – that is, not in trial projects, but in situations where the intervention has been funded on the basis of a relatively clearly articulated theory of change. Contribution analysis helps to confirm or revise a theory of change; it is not intended to be used to uncover and display a hitherto implicit or inexplicit theory of change. **The findings from a contribution analysis are not definitive proof, but rather provide evidence and a line of reasoning from which a plausible conclusion can be drawn that, within some level of confidence, the intervention has made an important contribution to the documented results.**

There are six standard steps that need to be taken to produce a credible contribution story:

1. **Establish the questions to be addressed.** Contribution analysis is less suitable for traditional causality questions such as 'Has the intervention caused the outcome?' 'To what extent, quantitatively, has the intervention caused the outcome?' These often are not that useful because they treat the intervention as a black box and do not get to the fact that there are usually many causes involved. Contribution analysis should be framed by questions such as 'Has the intervention influenced the observed result?' 'Has the intervention made an important contribution to the observed result?' 'Why has the result occurred?' 'What role did the intervention play? 'and for management questions: 'Is it reasonable to conclude that the intervention has made a difference?' 'What does the preponderance of evidence say about how well the intervention is making a difference?' 'What conditions are needed to make this type of intervention succeed?'
2. **Reconstruct the intervention logic.** Reconstruct the intervention logic describing how the intervention is supposed to work, which should lead to a plausible association between the activities of the intervention and the outcomes sought. The intervention logic must include the assumptions made and the inherent risks as well as external influences such as donor pressure, the influence of peers and resourcing levels. Some links in the intervention logic will be well understood or accepted. Others will be less well understood, contested or subject to significant influence other than from the intervention.
3. **Gather the existing evidence on the intervention logic.** It is useful to first use existing evidence such as from past related evaluations or research, and from prior monitoring, to test the intervention logic. What evidence (information from performance measures and evaluations) is currently available about the occurrence of the planned results? The links in the intervention logic also need to be assessed. What evidence currently exists on the assumptions and risks behind these links? Which are strong (good evidence available, strong logic or wide acceptance), and which are weak (little evidence available, weak logic or little agreement among stakeholders)? What evidence exists about the identified other influencing factors and the contribution they may be making?
4. **Assemble and assess the contribution story, or performance story, and challenges to it.** With this information, the evaluation team will be able to assemble the contribution story that explains why it is reasonable to assume that the actions of the intervention have contributed to the observed outcomes. The credibility of this story will need to be assessed – for example, would 'reasonable people' agree with the story? Does the pattern of results observed validate the results chain? Where are the main weaknesses in

the story? Weaknesses in the story point to where additional data or information is needed.

5. **Seek out additional evidence.** Having identified where the contribution story is less credible, additional evidence is now gathered by the evaluation team to strengthen the evidence in terms of what results have occurred, how reasonable the key assumptions are, and what has been the role of external influences and other contributing factors. This additional evidence can include the collection of new data such as from surveys, field visits, administrative data, focus groups, national statistical data etc. as well as the synthesis of evidence from other research and evaluations.
6. **Revise and, where the additional evidence permits, strengthen the contribution story.** Drawing on this additional evidence, a more substantive and credible story can be built, one that a reasonable person will be more likely to agree with. This does not mean that it is fool-proof, but the additional evidence will have made it stronger and more plausible.

Contribution analysis argues that a reasonable contribution to a given result can be made if the following pertain.

- **There is a reasoned intervention logic for the intervention.** The key assumptions behind why the intervention is expected to work make sense, are plausible, may be supported by evidence and/or existing research, and are agreed upon by at least some of the key players.
- **The activities of the intervention were implemented as set out in the intervention logic.**
- **The intervention logic – or key elements thereof –is supported by and confirmed by evidence on observed results and underlying assumptions.** The chain of expected results occurred. The intervention logic has not been disproved.
- **Other influencing factors have been assessed.** These are either shown not to have made a significant contribution or their relative role in contributing to the desired result has been recognised.

GLOBAL ELIMINATION METHODOLOGY

Scriven's GEM builds upon his earlier modus operandi (MO) method to provide an approach specifically geared towards substantiating causal claims. The approach entails systematically identifying and then ruling out alternative causal explanations of observed results. It is based on the idea that for any event it is possible to draw up lists of possible causes or alternative hypothetical explanations for an outcome of interest. Once this list of possible alternative explanations is drafted, data are collected and analysed to see which of the possible alternative explanations can be ruled out.

For example, take the case of the evaluation of an intervention providing farmers with improved seeds to increase their production and hence their income and well-being. If data collected by the evaluation team showed that the farmers had experienced an increase in their annual income, this might be because the intervention had been effective, or it might have been caused by something else. A list of possible causes might include the hypotheses that due to a drought in other areas, local farmers were able to get a higher price for their crops, even though they had not produced more. Or maybe their increased income had been from other sources such as hired labour. Data would then be collected and analysed to see if these possible alternative explanations could be ruled out. For example, if data about local prices showed they had been stable, increased prices as the reason for increased income could be ruled out. If data showed that income from hired labour had not increased over the period under review, it could also be ruled out as the cause of increased farmers' incomes and so on.

With GEM, each potential cause will have its own set of footprints or MO – a sequence of intermediate or concurrent events, a series of conditions or a chain of events that has to be present when the cause is effective (Scriven, 2008). GEM sets out to identify potential causes of effects by examining the facts of a case and establishing which MOs are present and which are not. Any cause for which the MO is not present can be dismissed, leaving only the causal explanations that have a genuine causal link. GEM

aims to provide a framework for evaluation which can establish causal claims beyond reasonable doubt.

REALIST EVALUATION

Realist evaluation seeks to understand how and why an intervention works differently in different contexts. The complete realist question is: 'What works, for whom, in what respects, to what extent, in what contexts and how?' Realist evaluators aim to identify the underlying generative mechanisms that explain **how** the outcomes were caused and the influence of the context. Based on specific theories, evaluation provides an alternative lens to empiricist evaluation techniques for the study and understanding of interventions. This technique assumes that knowledge is a social and historical product; thus the social and political context as well as theoretical mechanisms need consideration in analysis of intervention or policy effectiveness.

Realist evaluation techniques recognise that there are many interwoven variables operative at different levels in society. This evaluation method thus suits complex social interventions, rather than traditional cause-effect, non-contextual methods of analysis. The realist technique acknowledges that interventions and policy changes do not necessarily work for everyone, since people are different and are embedded in different contexts. Realist evaluation was popularised by the work of Ray Pawson and Nick Tilley in 1997. They described the procedure followed in the implementation of realist evaluation techniques in programme evaluation and emphasise that once hypotheses have been generated and data collected, the outcomes of the programme are explored, focusing on the groups that the intervention benefited and those who did not benefit. Interventions are viewed as open systems in which there are multiple and competing mechanisms that interact with the surrounding context to produce outcomes. Effectiveness of an intervention is thus not dependent on the outcomes alone (cause-effect); rather there is a consideration of the theoretical mechanisms that are applied, and the socio-historical context in which the programmes were implemented. To address the core question of what works, for whom, in what respects, to what extent, in what contexts and how, evaluators are asked to consider how underlying mechanisms are likely to interact with the historical and cultural context, location, economic and political structures etc. to produce varying outcomes. Evaluators consider the nature of a planned intervention, the target population and the context in which the intervention will operate to map out a series of hypothetical mini-theories of change called **context mechanism outcome configurations** which relate the various contexts of an intervention to the multiple mechanisms by which it might function to produce various outcomes. Realist evaluation draws on both quantitative and qualitative data sources to build a picture of the intervention in action and identify how mechanisms are operating in reality to revise, substantiate or invalidate hypothetical context mechanism outcomes.

QUALITATIVE COMPARATIVE ANALYSIS

Qualitative comparative analysis (QCA) systematically compares cases to identify clusters of factors that have produced the outcomes and impacts of interest. QCA is a means of analysing the causal contribution of different conditions (e.g. aspects of an intervention and the wider context) to an outcome of interest. QCA starts with the documentation of the different configurations of conditions associated with each case of an observed outcome. These are then subjected to a minimisation procedure that identifies the simplest set of conditions that can account for all the observed outcomes, as well as their absence.

The results are typically expressed in statements in ordinary language or as Boolean algebra, as the following example illustrates.

> A combination of condition A and condition B or a combination of condition C and condition D will lead to outcome E. In Boolean notation this is expressed more succinctly as A*B + C*D→E

QCA results are able to distinguish various complex forms of causation, including the following.

- **Configurations of causal conditions, not just single causes.** In the example above, there are two different causal configurations, each made up of two conditions.
- **Equifinality, where there is more than one way in which an outcome can happen.** In the

above example, each additional configuration represents a different causal pathway

- **Causal conditions which are necessary, sufficient, both or neither, plus more complex combinations.** These are known as INUS causes – insufficient but necessary parts of a configuration that is unnecessary but sufficient, which tend to be more common in everyday life. In the example above, no one condition was sufficient or necessary. But each condition is an INUS type cause.
- **Asymmetric causes, where the causes of failure may not simply be the absence of the cause of success.** In the example above, the configuration associated with the absence of E might have been one like this: A*B*X + C*D*X →e Here X condition was a sufficient and necessary blocking condition.
- **The relative influence of different individual conditions and causal configurations in a set of cases being examined.** In the example above, the first configuration may have been associated with 10 cases where the outcome was E, whereas the second might have been associated with only 5 cases. Configurations can be evaluated in terms of coverage (the percentage of cases they explain) and consistency (the extent to which a configuration is always associated with a given outcome).

QCA is able to use relatively small and simple data sets. There is no requirement to have enough cases to achieve statistical significance, although ideally there should be enough cases to potentially exhibit all the possible configurations. The latter depends on the number of conditions present. QCA is a theory-driven approach in that the choice of conditions being examined needs to be driven by a prior theory about what matters. The list of conditions may also be revised in the light of the results of the QCA analysis if some configurations are still shown as being associated with a mixture of outcomes. The coding of the presence/absence of a condition also requires an explicit view of that condition and when and where it can be considered present. Dichotomisation of quantitative measures about the incidence of a condition also needs to be carried out with an explicit rationale, and not on an arbitrary basis. Box 3.2.7 provides an example of how contribution analysis and QCA were used to answer a causal question in the midterm evaluation of the Africa Regional Empowerment and Accountability Programme.

BOX 3.2.7 Using contribution analysis design to answer a causal question

To address the subquestion 'Does evidence generation and citizen awareness about national, governance issues, and/or regional or continental issues support improvements/changes in national governance?', the midterm independent evaluation of the Africa Regional Empowerment and Accountability Programme (2015) used contribution analysis and QCA to analyse the data and determine the influence of the project on changes in national governance. The contribution analysis tested hypotheses and assumptions in the intervention logic about how activities were expected to produce outcomes, searching for evidence that supported and did not support the intervention logic. The contribution analysis also assessed alternative explanations for change to test the extent to which activities contributed to observed change. The QCA compared conditions present in successful versus unsuccessful interventions. The QCA was conducted to analyse the results of two of the evaluation questions to identify what internal and external conditions were necessary and sufficient for change to occur.

COMPARATIVE CASE STUDIES

Comparative case studies can be used to identify causal relationships between different variables. Mostly, cases are a way of looking at and comparing individual examples instead of a whole group. This is helpful when evaluators want to understand how and why something happens in detail, or to compare different cases. For example, it might be useful to compare a case where an intervention worked well with one where it did not to understand what made the difference. Or it might be desirable to look at a 'typical' case and an 'atypical' one to see what might be complicating factors.

There are different ways to select cases for an evaluation. Sometimes cases are chosen randomly; other times, they are selected based on certain

characteristics. Case studies can be used to understand how an intervention works, to compare different interventions, or to study typical and atypical outcomes. They can be selected with different purposes in mind:

- to bring to life what previously was an abstract or very simplified model of what is taking place;
- to help the evaluation team understand in detail the causal mechanisms connecting events which are correlated or associated;
- to identify other complicating and confounding influences that may require the intervention's theory of change to be modified;
- to verify the validity of the measures that had been used to describe the cases.

SEE: *Gerring and Cojocaru (2015).*

CASE ANALYSIS METHODS

The **within-case approach** involves looking at each individual case separately. This means looking at each person, family, organisation or other unit of analysis on its own. The **across-case approach** to qualitative data analysis involves looking at all the cases together for patterns that may emerge. Midway between the two approaches is the possibility of **case comparisons**. These typically take the form of two contrasting cases. Examples of these are cases with:

- most similar intervention design but different outcomes;
- most different intervention design but similar outcomes;
- typical intervention design and outcome versus atypical design and outcome.

PROCESS TRACING

Process tracing is another method used to examine and test a specific causal link in terms of whether the evidence is sufficient to draw a conclusion about the cause. It is a **case-based approach to causal inference** that focuses on the use of clues within a case (causal process observations) to adjudicate between alternative possible explanations.

Process tracing involves four types of causal tests:

- passing the **straw in the wind** test lends support for an explanation without definitively ruling it in or out;
- the **hoop** test is failed when examination of a case shows the presence of a necessary causal condition when the outcome of interest is not present;
- the **smoking gun** test is passed when examination of a case shows the presence of a sufficient causal condition – uncommon smoking gun conditions are more persuasive than common ones;
- the **doubly definitive** test is passed when examination of a case shows that a condition is both necessary and sufficient support for the explanation – this tends to be rare.

Process tracing can be used both to see if results are consistent with the intervention theory of change and to see if alternative explanations can be ruled out.

SUMMARY

All of the approaches outlined above aim to address attribution by examining the facts of a case to gain an in-depth understanding of the causal chain connecting observed outcomes to an intervention. Their goal is to explain what has occurred and how it has occurred. They either seek out evidence to substantiate whether an intervention's theory of change occurred in practice, or they do the same for a number of alternative causal hypotheses that outline what might have occurred if the causes or assumptions set out in the theory of change had varied. Evidence is gathered to assess each of the hypothesised explanations and to account for any external factors which may have played a role. Causation is established beyond reasonable doubt by collecting evidence to validate, invalidate or revise the hypothesised explanations, with the goal of documenting the links in the actual causal chain.

Using a combination of designs to answer causal questions produces an evaluation design with stronger evidence that the intervention has generated observed results, as the example in Box 3.2.8 demonstrates.

BOX 3.2.8 Using a combined design to answer a causal question

The retrospective impact evaluation of Save the Children's sponsorship programming in Ethiopia's Woliso Impact Area, 2002–2010 (Davidson and Chianca, 2020) used several causal inference strategies to attribute change to intervention activities; these included the following.

1. Using a hypothetical counterfactual
   - Using key informant interviews data to identify likely career pathways had the intervention not trained teachers and comparing the lifetime incomes of subsistence farmers to teachers
2. Identifying and ruling out alternative explanations for the observed changes
   - Whether other NGO or government support was (or was likely to have been) forthcoming over the period in question
   - Whether any of the changes made could have been initiated by communities themselves at the time (did they have capacity to make this happen?)
3. Checking for congruence of evidence with a causal relationship
   - Checking the timing of changes in relation to intervention activities to see whether they could plausibly have been influenced by Save the Children's work and/or other factors
   - Searching for disconfirming evidence and following up on exceptions (e.g. individuals or communities where the results were different, or much smaller or larger, to find out why)

3.2.5 Participatory approaches: overview

These approaches are distinguished from the theory-based ones discussed in the previous sections by the fact that they do not set out to address attribution of cause and effect as explicitly. In general, these **participatory approaches place stakeholder participation at the heart of data collection and analysis**. They target stakeholders such as target groups, final beneficiaries and implementers in order to establish which factors are perceived to have been important in producing change. In so doing, they aim to gain insights into how an intervention is performing and the role it is playing in bringing about change. Before looking at some of the more commonly used participatory approaches, the following provides an overview of what is understood by this type of approach.

ENGAGING WITH STAKEHOLDERS

Effective evaluation relies on stakeholder engagement. Normally, an evaluation involves various people and organisation types: donors, beneficiaries and other stakeholders. They can be representatives of public agencies (domestic or multilateral), non-governmental or private sector. Moreover, individuals and communities that are affected by an intervention can be directly interested in the evaluation process and its results. Ultimately, when an intervention is financed through public funding, evaluation is also a matter of delivering accountable results for the broader audience, notably the taxpayers.

The different stakeholders and audiences of evaluation activities are thus a very diverse group, with various interests and agendas. Sometimes they may even be in conflict. To satisfy their needs, evaluators often must invest significant effort to engage them in the actual evaluation tasks. There are various ways to do this, but the most important is to follow the principles of **participation and inclusivity**. Multiple evaluation methods and tools offer a wide range of possibilities to facilitate stakeholder engagement and meaningful contributions to the evaluation process.

Participatory evaluation is a term used to describe an evaluation that ensures engagement of evaluation stakeholders throughout the evaluation process. This involvement can occur at any stage of the evaluation process, from evaluation design to data collection and analysis and reporting of findings. The type and level of stakeholder involvement will necessarily vary between different types – for example, between a local-level impact evaluation and an evaluation of policy changes (Guijt 2014). It is important to consider the purpose of involving stakeholders, and which stakeholders should be involved and how, in order to maximise the effectiveness of

the approach. In this approach, it is assumed wide participation and inclusion of all relevant evaluation stakeholders will allow for improving the performance of evaluation and the intervention. By placing the emphasis on participation, it is important when utilising this approach to be clear on the intention of the evaluation. As for any evaluation, the aim is to understand what the primary users seek to know, and how information can be obtained which is useful to them. In participatory evaluation, the principle is the same, only the primary users of the evaluation are also its beneficiaries.

ADVANTAGES OF PARTICIPATORY EVALUATION

- Participatory evaluation, as it is co-designed by participants from within the community is naturally more relevant to reality, as are the evaluation questions. Thus, there is greater potential for ownership and far lower risk of findings and recommendations not being used.
- The process of conducting the evaluation is in itself developmental and encourages communities to think meaningfully about strategy and change. Through this, developmental evaluation can be empowering and can equip people with new skills. This process of discovery can also be useful for enhancing social cohesion, as individuals articulate their role and place in a team committed to creating social change.
- Participatory evaluation is considered pragmatic as the data, method of collecting data, and the story the data stand to tell are more relevant by virtue of the fact that those closest to the change define all these things.
- Often, this form of evaluation is able to explicate the hidden and even conflicting views or agendas of stakeholders and facilitate consensus negotiations.

DIFFICULTIES OF PARTICIPATORY EVALUATION

In general, participatory evaluation speaks to small, unsystematic and largely subjective data. Thus, it can be difficult to make broad statements about what worked and how to leverage change. Also, if this is facilitated, there can be a lack of consideration of context. Development practitioners may unwittingly bring in methods and approaches which simply do not fit with the system of values of certain communities.

SEE: *Better Evaluation website, Participatory evaluation web page; CIDA (2001); Community Tool Box, Section 6: Participatory Evaluation; UNICEF (2011).*

3.2.6 Commonly used participatory approaches

Several approaches have emerged in recent years in an attempt to better capture change processes based on more complex partnerships and with greater stakeholder involvement. Outcome mapping and social frameworks, for example, are better able to help partnerships describe, then monitor, what needs to be done to deliver the outcomes – and ultimately evaluate whether those outcomes have been or are being delivered. Participatory impact pathways, most significant change, qualitative impact protocol, outcome harvesting and the success case method are other approaches that have emerged in recent years in an attempt to gain better insights into how interventions are delivering.

OUTCOME MAPPING

Outcome mapping was originally developed by the International Development Research Centre in Canada, and the first comprehensive outcome mapping manual was produced in 2001. Outcome mapping seeks to identify and assess changes in the behaviour of people, groups and organisations with which an intervention works directly. It does not seek to prove causality or attribution for those changes, but instead attempts to show logical linkages between those changes and an intervention's activities, thereby enabling an intervention's contribution to change to be understood (Earl, Carden and Smutylo, 2001). Outcome mapping is a participatory planning approach, although it has implications for how monitoring and evaluation are conducted. It purposefully includes those implementing an intervention in both design and data collection to encourage ownership and the use of findings. It was designed

to be a 'consciousness-raising, consensus-building, and empowerment tool for those working within a development programme' (Earl, Carden and Smutylo, 2001, p. 4). Outcome mapping is designed to be used at the beginning of an intervention, after the main focus of that intervention has been decided.

There are three key stages to developing an outcome map (Earl, Carden and Smutylo, 2001).

- The first stage, **intentional design**, helps an intervention establish consensus on the changes it aims to help bring about and plan the strategies it will use. It helps answer four questions:
  - What is the vision to which the intervention wants to contribute?
  - Who are the intervention's boundary partners?

    NOTE: *Boundary partners are the individuals, groups or organisations with which the programme interacts directly and where there will be opportunities for influence. Boundary partners may be individual organisations but might also include multiple individuals, groups or organisations if a similar change is being sought across many different groups (e.g. research centres or women's NGOs).*

  - What are the changes that are being sought?
  - How will the programme contribute to the change process?
- The second stage, **outcome and performance monitoring**, provides a framework for the ongoing monitoring of the intervention's actions and the boundary partners' progress towards achievement of those outcomes. Monitoring is based largely on self-assessment.
- The third stage, **evaluation planning**, helps identify evaluation priorities and develop an evaluation plan.

Strengths and weaknesses

Unlike some advocates of the logframe, supporters of outcome mapping do not claim it appropriate in all situations. This, added to the fact that outcome mapping is rarely forced on organisations as a condition of funding, means debates surrounding outcome mapping are less intense than those surrounding the logframe. Some of outcome mapping's strengths for evaluation purposes can be summarised as follows:

- It introduces (monitoring and) evaluation at an early stage of an intervention and ensures that (monitoring and) evaluation is built into design.
- Because it is based on outcomes of observable behaviour change, it can be more intuitive for evaluators to grasp than the sometimes more abstract language of objectives.
- It encourages evaluators to assess both the outcomes of interventions – thus focusing clearly on change – and the processes through which those outcomes are generated.
- It is much better than linear evaluation tools at dealing with complexity. Outcome mapping does not seek to show direct attribution for change resulting from a single source. This means outcome mapping may be more appropriate for the evaluation of interventions with multiple inputs.
- Because of its focus on boundary partners, outcome mapping is good at dealing with interventions with a special focus on organisational change. It can therefore be used to support the (monitoring and) evaluation of capacity development – an area which people find particularly difficult to assess using more linear tools (see Simister and Smith, 2010).

However, outcome mapping is not appropriate in all circumstances. Some of its limitations have been described as follows:

- Because it deals with contribution rather than attribution, it cannot easily be used for processes that demand hard measurement of results, such as cost-benefit analysis and assessment of value for money.
- Outcome mapping may be best used at the level of medium-sized interventions. It is not a tool that is necessarily appropriate for handling large, complex ones, because it may be difficult to identify who will change and how. Earl, Carden and Smutylo (2001) point out that to be effective, outcome mapping must be sufficiently specific to enable the identification of key groups that will be influenced by an intervention. Equally, outcome mapping may not be appropriate for small interventions where

the investment of time needed would not be proportional to the likely benefits.

- The journalling approach to tracking progress means a lot of data are generated, creating challenges for data analysis.
- Outcome mapping is good at identifying changes within supported groups that are part of the process, partly because it encourages self-reflection and self-assessment. It may not be as useful at identifying change for people, organisations or groups that lie outside an intervention, such as the targets of policy influencing work.
- Outcome mapping does not focus predominantly on impact assessment. It recognises the need to look at long-term changes in people's lives brought about by development interventions but regards this as the responsibility of an intervention's boundary partners. If donors require in-depth impact assessment, outcome mapping needs to be supplemented by other tools and methodologies.
- In comparison with the logframe, outcome mapping is less able to provide a short, concise summary of an intervention.

Although outcome mapping can be, and frequently is, used as an approach in its entirety, it is often adapted, and can be used in conjunction with other methodologies such as the logical framework. Indeed, it is perfectly possible to embed an outcome map within a logframe, or set logframe indicators that can be generated by outcome mapping processes. Individual features of outcome mapping – such as the setting of progress markers at 'expect to see', 'like to see' and 'love to see' levels – are often used, even if the entire approach is not. Many organisations have also carried out work based on the principles of outcome mapping – such as participatory planning, understanding of complexity, valuing contribution rather than attribution – without necessarily adopting the process in its entirety.

NOTE: *The most comprehensive guide to outcome mapping is the guide* Outcome Mapping: Building Learning and Reflection into Development Programs *(Earl, Carden and Smutylo, 2001). There is also an outcome mapping community* website *which is regularly updated and contains much information on how outcome mapping is being used and applied. Further information, and a more comprehensive reading list, can be found at the* Better Evaluation *website.*

SOCIAL FRAMEWORKS

Social frameworks combine some of the principles of the logframe approach and outcome mapping. In social frameworks, the logframe's vertical logic of causal pathways is replaced with **expected pathways of influence** through networks of people, groups or organisations. While single pathways can be represented as a vertical sequence of events in a logframe matrix-type structure, representation of networks of expected influence requires the use of diagrams.

An actor-oriented approach, social frameworks draw on social network analysis methods and overlap with a number of outcome mapping practices. They differ from outcome mapping in that the delivery chain is traced along a series of actors from the end-user backwards, allowing decision makers to draw a pathway through the actor network to establish the responsibility each has for realising the intervention's intended outcomes. Social frameworks distinguish between different types of relationships in the network and, importantly, include those relationships which extend beyond the boundary partners. In doing this, they enable a more nuanced analysis of the nature of the challenges faced by the entire delivery network and allow a suite of progress markers to be developed which better reflect the path to the desired outcome.

Building a social network map helps assess the conditions needed to create an enabling policy environment among a diverse array of actors and to think clearly about the theory of change, focusing on the behaviour of actors rather than a disembodied set of outputs. Once a path has been established through a network, it is easier to clarify each actor's responsibility for delivering specific elements of the proposed change process. This then provides a platform to convene and engage the various boundary partners, allowing them to define their own progress markers – and provides a very useful framework for evaluators to assess that progress.

SEE: *Shaxson and Clench (2011).*

PARTICIPATORY IMPACT PATHWAYS ANALYSIS

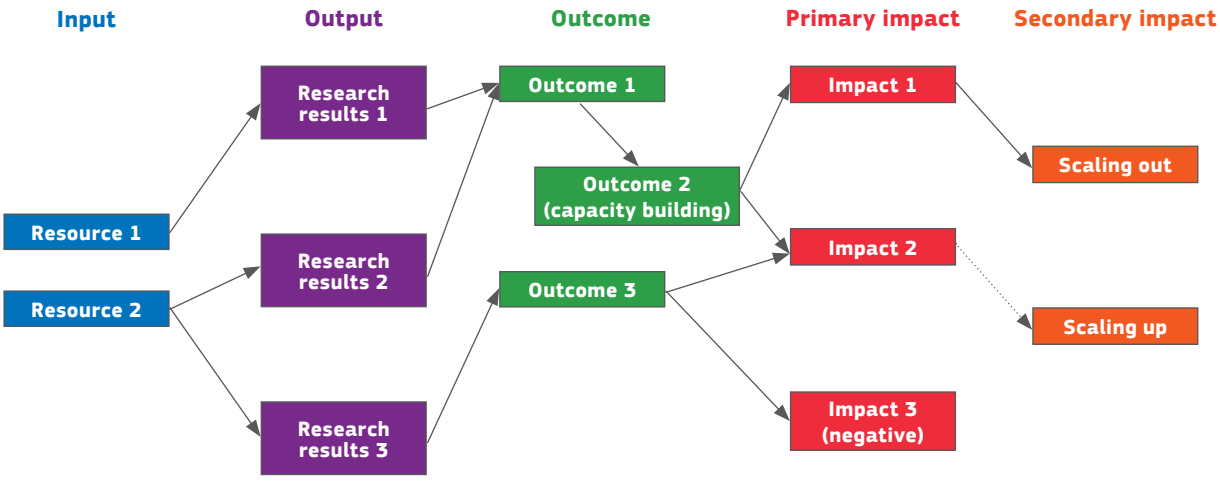
Participatory impact pathways analysis is similar in its philosophy to outcome mapping, with the main difference being that the former engages participants in predicting how outcomes can lead to social, economic and environmental impacts. Participatory impact pathways analysis involves stakeholders in **joint reflection on pathways leading to impact** to develop a theory of how a strategy, policy, instrument, modality or intervention can bring about the desired changes.

Participatory impact pathways analysis acknowledges that the **route from input to impact is not always straightforward**. For instance, one type of input may lead to different outputs and outcomes, which are also influenced by other inputs (see Figure 3.2.5). Impacts may be expected or unexpected, primary or secondary. This method is thus especially useful when evaluators attempt to look at impacts from a more complex perspective. Identification of multiple intervening factors and cumulative assessments are also feasible with participatory impact pathways analysis.

MOST SIGNIFICANT CHANGE

Most significant change (Davies and Dart, 2005) is a form of participatory monitoring and evaluation that involves the **collection and selection of stories of significant changes** that have occurred in the field. This approach answers questions about what is valued by different stakeholders (and also to generate data for developing and testing intervention logic). The central element of the most significant change approach involves the systematic collection and selection of a purposive sample of significant change stories. The stories themselves are elicited from intervention participants by asking them to relate what significant changes (positive or negative) have occurred in their lives in the recent past, and enquiring why they regard them as being significant. It involves collecting data in the form of stories about observed or experienced changes, and a systematic process to decide on the most significant stories by panels of designated stakeholders. The use of multiple levels of selection enables large numbers of significant change stories to be reduced to a smaller number of stories viewed as being most significant by different groups of stakeholders. The output of a small collection of verified performance stories can provide useful examples of what has been achieved by the intervention to complement other evidence.

FIGURE 3.2.5 **Impact pathway**



SOURCE: Adapted from CIRAD (2015), based on Douthwaite et al. (2003).
NOTE: ⟶ main contribution; ⋯▸ partial contribution.

Similar approaches to most significant change include the success case method, appreciative inquiry and positive deviance.

- The success case method deliberately looks at the most, and least, successful participants of an intervention. The purpose is not to examine the average performance. By identifying and examining the extreme cases, it provides information about what the intervention produces when it works well. It is suitable for interventions where a small number of highly successful cases would be enough to justify the investment in an intervention (e.g. investment in economic development interventions) and where what is learned might improve value for money by improving targeting.
- The appreciative inquiry approach focuses on identifying instances of 'peak performance' with those involved in an intervention or an organisation, analysing how it came about and what might be done to support more of this.
- The positive deviance approach also works by identifying and learning from successful outliers, in this case exceptional cases or 'positive deviants'. A particular feature of this approach is that those whose behaviour is intended to be informed and changed by the evaluation are involved in the process of identifying the outliers, gathering information about them to understand how they can achieve exceptional results, and then developing recommendations for their own behaviour that learns from these.

SEE: Most Significant Change (MSC) *web page on Rick Davies's Monitoring and Evaluation NEWS news service.*

QUALITATIVE IMPACT PROTOCOL

Qualitative impact protocol (QuIP) provides an independent reality check of the intervention logic, gathering evidence of a intervention's impact through narrative causal statements collected directly from purposefully sampled intended intervention beneficiaries. Respondents are asked to talk about the main changes in their lives over a pre-defined recall period and prompted to share what they perceive to be the main drivers of these changes, and to whom or what they attribute any change – which may well be from multiple sources. Importantly, the interviewers are blind to the intervention being evaluated to reduce confirmation bias in their data collection.

OUTCOME HARVESTING

Outcome harvesting is mainly used to retrospectively identify emergent outcomes by collecting evidence of what has changed and working backwards to identify whether and how an intervention has contributed to these changes. This approach was developed to be able to report on the achievements of complex, emergent interventions, such as funding a network or community development, where the specific outcomes and the pathways to them will be emergent in response to opportunities, and it is not possible to develop a detailed intervention logic in advance to inform the evaluation. It can also be used for evaluations with learning objectives and monitoring purposes in contexts where the intervention logic is not or cannot be sufficiently defined.

SEE: *The Safer World Learning Paper to understand how outcome harvesting could be adapted to use for different purposes, such as monitoring in conflict situations.*

SUMMARY

As with the theory-based approaches, these participatory ones gather information that can help reconstruct the intermediate steps between cause and effect. However, they do not make causal explanation their primary goal. In practice, these **approaches are not necessarily intended to be stand-alone exercises but can instead be usefully employed as one element within a wider evaluation framework**. Using a combination of designs to answer causal questions produces an evaluation design with stronger evidence that the intervention has generated observed results.

3.2.7 Other approaches

There are other types of evaluation approaches which are neither theory based nor participatory. These include for example, a systems approach and multi-layered models.

SYSTEMS APPROACH

A systems approach, based on systems theory, assumes that a strategy, policy, instrument, modality or intervention **occurs within a larger complex system consisting of interconnected elements**. A systems view incorporates a much wider perspective than does a typical logframe, encompassing all the surrounding events and actors that could affect the planned outcomes. It provides quantitative methodologies that enable evaluators to consider the dynamic relationships of factors at multiple levels of analysis, but it also includes qualitative approaches to actively engage members of the community in a participatory process. This information is usually conveyed in a network diagram like Figure 3.2.4, often called a causal loop diagram because of its many different types of feedback loops. Advocates of a systems approach emphasise three aspects of these representations:

- boundaries – what is excluded from the model;
- the structure of relationships;
- the particular perspective applied – the kinds of actors, events and relationships diagrammed.

SEE: *Fujita (2010).*

MULTI-LAYERED MODELS

Multi-layered models offer a way to manage some of the problems associated with complexity, notably the ability to **include sufficient detail without making the representation so complex it cannot be understood.** This detail is provided through a process of nesting, whereby a simple large-scale model (diagram) is associated with smaller-scale models describing specific aspects of the larger model. If necessary, each of these can in turn be supported by even smaller, more detailed sub-models. Multi-layered models can be developed using dedicated software or multiple web pages connected with hypertext links.

NOTE: *The Donor Committee for Enterprise Development (DCED) 'Evidence Framework' organises robust research on results in private sector development based on the logic by which interventions typically expect to achieve pro-poor impacts. It is designed as a 'clickable' results chain that signposts key evidence for each step in the logic. See:* DoView *for dedicated software; and the DCED* Evidence Framework *for an example of a nested set of hyperlinked web pages.*

SECTION 3.3

Data collection and management

Meaningful evaluation depends on a good methodological design, tailored to the specifics of the evaluand in order to meet the requirements and intended uses of the evaluation.

The evaluation methodology provides the detailed framework that will allow the evaluators to answer the evaluation questions and arrive at an overall assessment of the intervention. In addition to the agreed-upon evaluation questions, judgement criteria, indicators and targets, the evaluation methodology:

- documents the combination of tools that will be used to collect data – for example, key informant interviews, documentation, surveys, polls, photographs;
- ensures the validity of the findings – in particular:
  - the timing of data collection;
  - sampling choices;
  - the analysis to be carried out to make sense of data;
  - triangulation – best available evidence drawn from a diverse and appropriate range of methods and sources (EC, 2021a).

This section starts with an overview of the links between data, information and knowledge, as well as presenting the concepts of quantitative and qualitative and primary and secondary data. It then goes on to briefly describe some of the more commonly used data collection tools – traditional as well as emerging ones driven by information and communication technologies (ICTs) and sampling methods. It addresses the issue of data management, including data architecture, quality assurance and data visualisation.

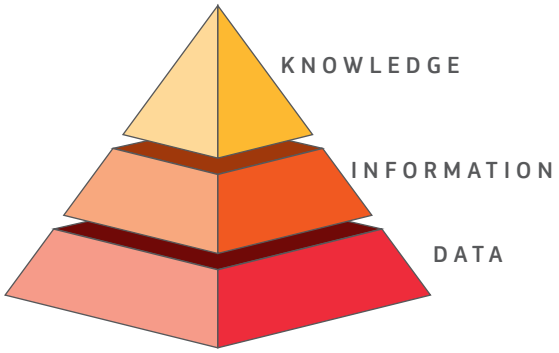
3.3.1 Data, information and knowledge

One of the main challenges of evaluation is to present a comprehensive picture of the complexity associated with the evaluand. To this end, evaluators collect data and, based on these data, generate information and knowledge. These crucial components of evaluation are intended to inform decisions regarding implementation, financing and policymaking.

- **Data** are the raw facts and figures collected through data collection methods such as surveys, interviews and observation. They are a collection of discrete values that convey information, describing quantity, quality, facts, statistics and other basic units of meaning. Data can come in the form of text, observations, figures, images, numbers, graphs or symbols.
- **Information** is data that have been organised and presented in a meaningful way, often through data analysis: the process of organising, exploring and making sense of the data (which is explored in Section 3.4). Information refers to facts based on evidence. In evaluation, this evidence is supplied through the collected data.
- **Knowledge** is information that has been interpreted and synthesised to form a new understanding. It is the result of applying data interpretation and analysis to create new insights. It is gained through growing familiarity with data and information generated via the evaluation process. This results in an understanding of why and how things happened, why this way and not another way, and what the most optimal solutions are to given problems. In evaluation, knowledge is closely associated with the learning process, to improve aspects of the evaluand and minimise the risk of failure. Knowledge is necessary for informing decisions. It can be based on expert judgements or involve citizens at large (knowledge sharing, crowdsourcing).

Figure 3.3.1 presents a visualisation of these elements.

FIGURE 3.3.1 **The knowledge hierarchy**



QUANTITATIVE AND QUALITATIVE

Evaluation practice distinguishes between qualitative and quantitative data (see Figure 3.3.2 and Table 3.3.1). Both types of data can be collected separately or simultaneously, depending on evaluation requirements. The two types of data complement each other and bring a better, fuller picture of the evaluand, its results, outcomes and impacts.

- **Quantitative data** are principally intended to provide numeric answers; they respond to questions such as 'How many?' and 'How much?' These data can be gathered in a variety of ways, including both traditional and emerging methods. Quantitative data are especially useful for situating information on a temporal scale (before, during and after implementation), allowing for simple comparisons of what has changed as a result of the evaluand and to what extent. One of the main advantages of quantitative data is that numbers are easy to generalise and understandable for broader stakeholder groups. However, quantitative data can sometimes be difficult to obtain, incomplete or not sufficiently robust; in more complex cases, it may require an advanced knowledge of statistics and computational methods.
- **Qualitative data** characterise an object or event, and are non-numerical. They are also referred to as categorical data, since they can be organised as a set of properties. Whereas qualitative data have traditionally been collected through focus groups, interviews and observation, ICTs present emerging opportunities, such as participatory

FIGURE 3.3.2 Quantitative versus qualitative data collection

TABLE 3.3.1 Quantitative and qualitative data

Data type	Typical evaluation questions	Advantages	Disadvantages
Quantitative	How many? How much?	• Provides generalisable results due to larger sample sizes • Facilitates comparisons	• Difficulty in capturing complex phenomena • Limited depth and flexibility
Qualitative	Why? How?	• Provides rich, in-depth insights • Facilitates understanding of context	Limited generalisability due to smaller sample sizes

videos and photography/satellite images. Among the advantages of qualitative data in evaluation is the opportunity they afford in gaining a richer understanding of an object or event, which can help explain why and how something happened because of actions of the evaluand.

Thus, if an evaluation is looking at the impact of an intervention on employment rates, quantitative data would be used to measure, for example, the percentage of participants who found employment after completing the training offered by the intervention. Qualitative data would be used, for example, to understand the experiences of participants and why they think they succeeded (or failed) to find employment.

Data can be further characterised as primary or secondary.

- **Primary data** are original data, newly collected by the evaluation team for the purposes of evaluation, and could come from data collection tools such as surveys, interviews, focus groups, observation etc.
- **Secondary data** already exist and are not collected by the evaluation team; these data can come from published sources such as newspapers, magazines, books, databases, government reports etc.

Usually, **primary data collection** requires an additional investment of time and money – and in many cases, also entails travel and fieldwork. In contrast, **secondary data** are generally more readily available and are often free of charge. However, such

data may be less detailed and are often less tailored to the specific evaluation purposes. Increasingly, secondary data are available in public databases and platforms in an open-access format. Box 3.3.1 lists some of the most relevant sources of secondary data for international cooperation.

BOX 3.3.1 Free, publicly available databases relevant to cooperation

The following list, which is by no means exhaustive, presents links to popular publicly available databases covering social, economic, environmental and other issues relevant to cooperation interventions. Depending on the database, it is possible to find records for individual countries or regions or aggregated in other dimensions.

- European Commission - Joint Research Centre Africa Knowledge Platform
- EU's official data portal, data.europe.eu
- EU Science Hub's Joint Research Centre Data Catalogue
- EU's Eurostat database
- European Space Agency's Copernicus data sets
- EU's European Spatial Data Infrastructure (Inspire) Geoportal
- European Environmental Agency's data and maps
- Global Climate Monitor
- Global INFORM Risk Index
- Natural Earth public domain map data set
- Open Street Map
- OECD.stat data sets
- Food and Agriculture Organization of the United Nations (FAO) FAOSTAT food and agriculture data
- United Nations Development Programme (UNDP) Human Development Data
- United Nations Environment Programme (UNEP) Open Data
- UN Environment Programme World Conservation Monitoring Centre (UNEP-WCMC)
- UN Women, Women Count data dashboards
- US National Air and Space Administration (NASA) open data portal, data.nasa.gov
- US Geological Survey EarthExplorer
- US National Oceanic and Atmospheric Administration (NOAA) Open Data Dissemination
- World Bank Open Data

3.3.2 Data collection methods and tools

Data, both quantitative and qualitative, can be collected either in situ (through field activities – primary data) or ex situ (through desk activities – secondary data), and by using various tools, ranging from more traditional ones such as document reviews and interviews to some of the newer, ICT-driven ones such as big data.

TRADITIONAL

Traditional tools – those that have been in use for a long period of time – include the following.

- **Document reviews.** Documents directly or indirectly related to an intervention will be a key source of information for all evaluation teams. One of the challenges facing an evaluation team will be to identify the most relevant/useful documents from the typically vast range available.
- **Questionnaires/surveys.** These collect data from specific groups by means of a set of questions, logically connected with the evaluation questions, judgement criteria, and indicators and can be closed or open ended depending on evaluation needs.
  - **Closed-ended questions.** Responses are limited to a pre-set choice, such as a binary 'yes/no' or multiple choice, where more options are given. Other options for this type of question involve using a rating scale (e.g. Likert scale, semantic differential scale, rank-order scale).
  - **Open-ended questions.** These are free-form questions that allow respondents to answer in open-text format based on their complete knowledge, feelings and understanding. The response to such questions is not limited to a set of options. Unlike a closed-ended question, that leaves survey responses limited to the given options, an open-ended question allows evaluators to probe deeper into the respondent's answers, gaining valuable information about

the issue. The responses to these questions can be used to attain detailed and descriptive information on a topic.

A common error with questionnaires and surveys is to include too many questions, which runs the risk of a low response rate. As is the case when identifying the evaluation questions themselves, care should be taken to limit the number of questions in a survey/questionnaire to the bare minimum.

NOTE: *Read more about* How to Conduct Surveys *on Capacity4dev's Evaluation methodological approach wiki.*

- **Interviews.** These differ from surveys due to the possibility of reiteration that they offer: focusing in on particular responses with further clarifying questions. Information gained from interviews is not limited to a predetermined set of questions. Rather, the interviewer can direct the conversation in the direction he/she chooses based on the responses of interviewees. Interviews are especially useful for gaining a deeper understanding of issues but can be time-consuming and difficult to interpret as interviewees may be biased in their views. Different types of interviews are used in evaluation:
  - **Key informant interviews.** These are usually qualitative and aimed at gaining an in-depth perspective from people considered most informed about the specific evaluation topic.
  - **Structured/unstructured interviews.** In the first type of interview, questions are specifically tailored and seek precise information. In the second case, questions are broader and more open. Their relative advantages and disadvantages mirror those of the open-ended and closed-ended questions in surveys.
- **Oral histories.** These enable data collection through recorded interviews with individuals. Using this tool, specific narratives are examined, in order to obtain data needed for the evaluation tasks. Oral histories are usually recorded in audiovisual format or written transcriptions, where individuals share their accounts of family life, significant events, memories from the past and other topics of importance.

NOTE: *Read more about carrying out* Interviews *on Capacity4dev's Evaluation methodological approach wiki.*

- **Observation.** This refers to data collected while studying behaviour (e.g. of target groups) or things (e.g. infrastructure). There are different types of observation that can be used, varying in the degree of direct involvement of the evaluator. An evaluator can be either visible or invisible to the participants. Observation allows for in-depth, hands-on knowledge which is more valid/trustworthy than that gained from one-off interviews but can be very time-consuming (and costly). The choice of the right data collection method through observation should be made carefully, taking into account both methodological challenges and ethical aspects – for example, people under observation may behave differently or may experience unease or fatigue and can under- or over-perform their tasks (Hawthorne effect).
- **Focus groups.** A focus group is a group interview involving a small number of demographically similar people or participants who have other common traits/experiences. Their reactions to specific evaluator-posed questions are studied. Focus groups are used to better understand people's reactions to products or services and their perceptions of shared experiences. The discussions can be guided or open. As an evaluation tool, they can elicit lessons learned and recommendations for performance improvement. If group members are representative of a larger population, those reactions may be expected to reflect the views of that larger population.

  Focus groups constitute an evaluation tool that evaluators organise to collect qualitative data through interactive and directed discussions. Group members are often free to talk and interact with each other. Instead of an evaluator asking group members questions individually, focus groups use group interaction to explore and clarify participants' beliefs, opinions and views. The interactivity of focus groups allows evaluators to obtain qualitative data from multiple participants, often making focus groups a relatively expedient, convenient and efficacious tool. More advanced techniques for organising focus groups make use of interactive

facilitation methods, such as group exercises, and are supported with audiovisual material. Focus groups are a relatively cost-efficient means of data collection; however, in order to succeed, they require a skilled facilitator.

- **Participatory rural appraisal.** This data collection tool is frequently used in the context of rural and marginalised communities, often characterised by low levels of literacy. The basic techniques used include:
  - understanding group dynamics – for example, through learning contracts, role reversals, feedback sessions;
  - surveying and sampling – for example, transect walks, wealth ranking, social mapping;
  - interviewing – for example, focus group discussions, semi-structured interviews, triangulation;
  - community mapping – for example, Venn diagrams, matrix scoring, ecograms, timelines.

  To ensure that people are not excluded from participation, these techniques avoid writing wherever possible, relying instead on the tools of oral and visual communication such as pictures, symbols, physical objects and group memory. This tool serves not only data collection purposes but may also play a role in empowerment of citizens, women, children and others.

- **Participatory action research (also participatory learning in action)** builds upon engaging stakeholders in the evaluation process. This method of data collection is iterative; that is, it involves a set of events where stakeholders are invited to exchange and contribute to the evaluation process. In this way, data can be obtained and validated in a cycle of events, which simultaneously serves the purposes of improving the intervention delivery and enhancing learning from its successes and failures. The process often results in a greater awareness of stakeholders about the intervention and co-ownership of evaluation results. In participatory action research, stakeholders are no longer mere objects of evaluation, but equally contribute to the generation of data as co-evaluators. An example would be setting up a community of practice that deals with a particular intervention challenge. The community of practice interacts regularly and delivers data on the relevant topics (action), which are also combined with learning sessions.

In all cases, evaluation teams need to be aware of any potential biases in the data they collect using these traditional tools. Some typically occurring biases are summarised in Box 3.3.2.

BOX 3.3.2 Sources of bias in data collection

Evaluation team members should be constantly aware of potential biases such as the following.

- **Confirmation bias** – the tendency to seek out evidence that is consistent with the expected effects instead of being open to receiving evidence that could disprove them
- **Empathy bias** – the tendency to create a friendly (empathetic) atmosphere, for example, for the sake of achieving a high rate of answers and speedy completion of interviews, with the consequence that interviewees make overly positive statements about the intervention.
- **Self-censorship** – the reluctance of interviewees to freely express themselves and to depart from the views of their institution or hierarchy, simply because they feel at risk or uncomfortable sharing their opinions freely.
- **Strategy of interviewees** – purposely distorted statements with a view to influence evaluation conclusions in line with their own views.
- **Question-induced answers** – answers that are distorted by the way questions are asked or the interviewer's reaction to answers.

The evaluation team can reduce the potential impact of these biases and improve the reliability of data by:

- asking open questions, which prevents confirmation bias;
- mixing positive and negative questions, which prevents empathy bias and question bias;
- constantly focusing on facts, which allows for subsequent cross-checking of data and prevents interviewees' strategy bias;
- promising anonymity (and keeping that promise), which prevents interviewees' self-censorship.

EMERGING AND ICT-DRIVEN

In recent decades, evaluation has been supported with new and often ground-breaking data collection tools leveraging ICTs. These tools often allow data to be sourced more quickly and cost-effectively than had previously been the case. Moreover, available data have become ever more abundant, leading to data deluge – a situation where the available data are excessive and require an intensive effort to be processed. Although emerging and ICT-driven tools are new and appealing, they may carry unassessed risks.

> **SEE:** *Hassnain (2020) for more about mitigating these risks.*

- **Big data** is a 'hot' topic in evaluation practice. It means the use of data in large amounts, which is often characterised by a high degree of complexity. Typically, big data are generated online, or automated, and rely on virtual interactions between people. Increasingly, they are also available in real time, where the information technology (IT) system allows for constant updating. Examples include information about road traffic in Google maps or tweets updating with hashtags on Twitter. This type of data may directly contribute with numbers to an evaluation but also provides an additional picture of the situation, in which changes occur. Data derived from social media, for example, can help measure public perception (interest or trust of citizens) of an intervention.
  - **Social media** are data collection tools relying on popular websites and applications (e.g. Facebook, YouTube, Instagram, Twitter, LinkedIn). Social media enable the creation and sharing of content by users, which can offer rich resources for evaluators. Moreover, some apps and websites have embedded monitoring tools, which generate additional data about their users and web traffic. An example of popular tracing software is Google Analytics, which delivers detailed data on the use of a given website. Social media are especially advantageous for accessing contemporary attitudes and sentiments of relevant groups. They are however less reliable in terms of sampling, often relying on a sample that is unknown or unrepresentative. There are also significant risks that social media can be contaminated by bots or external agents.
  - **Crowdsourcing** is another popular forum for virtual interactions between different parties engaging in an evaluation. Contemporary crowdsourcing often involves digital platforms to attract and divide work between participants to achieve a cumulative result. The advantages of using crowdsourcing include lowered costs, improved speed, improved quality, increased flexibility and/or increased scalability of the work, as well as promoting diversity and inclusion.

    > **NOTE:** *A good example of a crowdsourced platform is the Global Forest Watch, where thousands of users contribute and use forest-related data for various purposes. This platform builds extensively on the use of geographic information systems (GIS) and related methodologies.*

- **Social network analysis** allows for obtaining data on social structures relevant for the intervention. This method of data collection relies strongly on network and graph theories. Social structures that are typically evaluated with social network analysis include stakeholder groups and their relationships. Changes in their interactions and network structures can be monitored over time. Such analysis can also serve as a diagnostic tool when difficulties occur, as it enables a quick identification of the people and groups causing them. The collection of data with social network analysis can be done in a traditional way (e.g. via sociometry, from which it originates) by simply posing questions to participants in a workshop. More complex interactions can be evaluated with the help of big data, mobile and geospatial methods (e.g. using a COVID-19 application to trace the density of human contacts in a given location). For analytical and visualisation purposes, an evaluator needs to become familiar with the relevant software (e.g. UCINET, Gephi, NetworkX, NetMiner, R).
- **Geospatial tools** are another innovation in the evaluation field. They are applicable where an intervention can be precisely defined and measured within a geographical scope and time

frame. Geospatial tools are a rapidly developing field with many applications, including data collection, analysis and communication. They can be practical in a variety of intervention sizes and sectors. While geospatial methods are diverse, they typically rely on a **geographic information system (GIS)**, which is a specific framework for gathering, organising and analysing data.

Using GIS, evaluators can gain insights into the relationships and patterns within geospatial dimensions. Geospatial methods can also offer an attractive alternative to randomised control trials. Using spatial location, layers of information can be generated and processed into visualisations (maps).

Geospatial data and tools supported with GIS typically include spatially explicit intervention data – available geocoded data with coordinates of the latitude and longitude of a given intervention. Another type of geospatial data is spatially explicit outcome and covariate data – that is, the georeferenced data fused with in situ or remotely sensed data describing outcomes and covariates. Thanks to the spatial data infrastructure, more advanced applications of GIS-generated data enable joining the intervention, outcome and covariate data into a common unit of observation. Further opportunities exist for using econometric tools that account for the unique features of spatial data (BenYishay et al., 2017).

- **Participatory GIS** is another tool to collect valuable georeferenced data. It typically combines participatory learning-in-action with the GIS. Stakeholders can be involved at each stage of the intervention, including evaluation tasks. Participatory GIS uses a similar logic as crowdsourcing. Data are sought from people who are directly concerned with the intervention including the end users. The participatory GIS toolbox normally requires strong communication components, such as satellite imagery and aerial photography. Participants may use more traditional forms of data collection (e.g. drawing a map) or innovative ICTs (e.g. a smart screen on which maps are displayed and can be moved by touch). This form of evaluation data collection can equally serve learning purposes, discussions and advocacy.
- **Online surveys** are increasingly replacing traditional paper-based questionnaires. They are relatively cost-efficient and quick in reaching respondents who possess at least basic internet literacy. Internet access is the main condition that needs to be met (including for software like Google, SurveyMonkey, Mentimeter). Surveys can be distributed via direct mail or on websites and social media (e.g. voting on Instagram). They are applicable at the different stages in the evaluation process and can be very useful, especially for measuring levels of satisfaction of intervention beneficiaries. An advantage of this tool is usually associated with the pre-structured means of viewing the aggregated data and simplifying the early stages of any analysis. However, this tool may also diminish the willingness or interest of people to participate in online surveys. Repeated telephone calls or other forms of follow-up may be required to ensure adequate levels of response.
- **Mobile phones, USSD and SMS** are used to collect data in different ways. Respondents can be asked to participate in a survey using an SMS service ('traditional' SMS or applications such as WhatsApp and Messenger). Data can also be collected through tracing of phone and application usage and Global Positioning System (GPS). The latter source is, however, severely constrained by legal provisions (data protection) and the practices of commercial companies managing the relevant messaging systems. An advantage of this tool lies in its ability to reach large numbers of people quickly. However, it may be limited to gathering relatively simple items of information.
- **Participatory videos and digital photography** allow for the collection of data with the help of audiovisual narratives. Using this tool, an evaluator facilitates interaction with stakeholders who jointly develop the script, aimed at filming selected aspects of an intervention. A camera is used that can be either professional grade or from a mobile phone. A community of stakeholders creates their own film that delivers insights for the evaluation practice. Further steps are showing the footage on the screen and joint editing. Similar to participatory action research, this way of data collection serves community empowerment purposes.

These less traditional methods of data collection are particularly relevant for evaluations in fragile contexts where the capacities of evaluators and stakeholders to access relevant data sources and respondents is more constrained due to travel limitations or security issues (Hassnain, 2019; Hoogeveen and Pape, 2020). This issue is further discussed in Subsection 3.3.5.

NOTE: *Useful resources are OECD: States of Fragility Report (2016); Hassnain (2020); Using Using ICTs in Evaluations in Fragility, Conflict and Violence (PowerPoint presentation); Hoogeveen and Pape (2020); and Hassnain, Kelly and Somma (2021).*

3.3.3 Sampling

Sampling is one the key activities with which evaluators need to be familiar, because it is generally not feasible for them to reach all the people/sites targeted by an intervention. While an intervention is targeted at a specific population, a sample is a smaller unit of that bigger whole, which represents certain population characteristics, and which can be studied in order to draw conclusions about the success or failure of an intervention. Unlike a census, which is typically performed by public statistical agencies and includes all members of the population, sampling covers only a select group of individuals.

Sample size determination is an important element of the sampling strategy and depends on the evaluation needs and resources. The credibility of an evaluation, especially its quantitative aspects, is strongly dependent on the number of cases in a dedicated sample. In small populations, the sample may simply cover all members. When the evaluated population is large, sampling is necessary as it would be difficult or too costly to reach each population member directly. In experimental design, sample groups may differ in size from each other. They can still, however, be compared with the help of appropriate statistical methods. Nevertheless, determining an adequate sample size is very important for reducing errors in statistical hypotheses' testing and minimising the confidence intervals.

Sample size determination is a complex process and may require statistical expertise or the use of software to perform power calculations. When determining sample size, the research question, population size, research design, statistical analysis and available resources should be carefully considered. Power analysis should be used to estimate the necessary sample size based on the desired level of statistical power, the expected effect size and the significance level. A minimum of 80 per cent power is generally recommended.

NOTE: *Dedicated software and sample size calculators can be helpful in determining the right size; see, for example, SurveyMonkey Sample Size Calculator and Creative Research Systems Sample Size Calculator.*

There are different sampling strategies, dependent on the context and evaluation needs. Their main characteristics are described below.

PROBABILITY SAMPLING

A probability sampling method is any method of sampling that utilises some form of random selection. In order to have a random selection method, a process or procedure needs to be set up that ensures that the different units of the population have equal probabilities of being chosen. Random samples become more representative as the population size increases. In small populations, random samples can be unrepresentative.

The most critical requirement of probability sampling is that everyone/every unit in the population has a known and equal chance of being selected, such that they are a fair representation of the population as a whole. This allows for generalisation of the results to the entire population – that is, external validity. In evaluation, two main approaches to probability sampling are commonly used.

- **Simple random sampling.** This approach involves a simple lottery, where all members of a given population participate and can possibly join a sample group. In case of a very large population, computer software can assist in this task or a random numbers book can be used.
- **Stratified sampling.** Members of a given population are grouped according to similar characteristics (e.g. sex, age, nationality); a simple

random sampling is used to extract representative proportions of the population from each of the groups.

The main benefit of stratified random sampling is that it can help to ensure that each stratum in a population is represented proportionately in the sample. This can be important if there is significant variation within the population and there is a need to ensure that this variation is reflected in the sample. Simple random sampling can sometimes result in a sample that is not representative of the population, especially if the population is large and/or heterogeneous.

Sometimes, however, the features of a given population are very complex and require advanced calculations in order to arrive at the right (representative) sample. While probability sampling is very valuable from the perspective of statistics and the credibility of results, in such cases, it can be very time-consuming to extract the right sample for the purposes of an evaluation.

NON-PROBABILITY SAMPLING

This kind of sampling is used when evaluation resources or knowledge about the population are limited. When randomness of the sample cannot be guaranteed, some population members will have a greater chance of being selected for the evaluation sample than others. Non-probability sampling approaches include the following; these are summarised in Table 3.3.2.

- **Snowball sampling** is based on a chain of referrals. Respondents who participate in the evaluation are asked to recommend other possible participants who can help with the collection of relevant data and information.
- In **convenience sampling**, little determines the sample size and characteristics. The only criterion is the actual availability of participants for the purposes of the evaluation. Evaluators may determine the choice of participating sample units.
- In **quota sampling** (also called **cluster sampling**), the population is broken down into specific groups according to their features (e.g. age, location, profession). A required number of participants is identified for each group (a quota). If possible, when designing the quotas, the evaluator should take into account how common each group is in the population as a whole.
- In **self-selection sampling**, people are invited to volunteer as respondents of the sample. This can be done through an official call for applications or other forms of invitation. While this may be

TABLE 3.3.2 Approaches to non-probability sampling

Method	Pros	Cons
Snowball sampling	• Can be used to reach hard-to-reach populations • Does not require a complete list of members of the population	• May lead to bias if the population is not homogeneous • May be time-consuming
Convenience sampling	• Easy to implement • Does not require a complete list of members of the population	• May lead to bias if the population is not homogeneous • May not be representative of the population
Quota sampling	• Can be used when a complete list of the population is unavailable • May be less expensive than other methods	• May lead to bias if the population is not homogeneous • May not be representative of the population
Self-selection sampling	• Does not require a complete list of members of the population • May be less expensive than other methods	• May lead to bias if the population is not homogeneous • May not be representative of the population
Purposive sampling	• Can be used to reach hard-to-reach populations • Does not require a complete list of members of the population	• May lead to bias if the population is not homogeneous • May be time-consuming

the easiest process to administer, the evaluator is unaware of the likely responses of those who did not self-select.

- In **purposive sampling** (also called **judgement sampling**), the evaluator is allowed to decide on the sample units. This is mainly used when a population is very small, or units of interest are characterised by a very rare occurrence. For example, evaluators may choose respondents with special characteristics they trust could be most suited to answer the evaluation questions. The key requirement here, for any form of replicability and thus credibility, is transparency in how the sample is chosen.

SEE: *Pell Institute,* Evaluation Toolkit*; Better Evaluation website,* Sample *web page.*

3.3.4 Data management

Data management is an increasingly important element of successful evaluation processes. It is important to plan ahead, to ensure data are stored in the most appropriate format and to be mindful of data quality assurance. Visualisation of data is also essential in order to effectively communicate the complexity of an evaluand to stakeholders. This subsection explores the different components of data management, quality assurance and visualisation.

DATA ARCHITECTURE

Data architecture comprises models, rules and standards that structure the way data is managed, including data collection, storage and integration. At the start of the evaluation process, evaluators should be clear on what this architecture will look like and possibly involve IT specialists and other relevant personnel in the planning. Reflecting on data architecture can be especially useful for interventions that are complex and require significant data volumes to be processed. The most important elements of the data architecture are described in the following paragraphs.

An evaluator should be able to identify how data will be **stored** for immediate use during the evaluation process and over the longer term. A storage medium (or media) needs to be designated. This can be, for instance, a personal computer (PC) or a web-based cloud, which are normally well suited to store data in electronic format (such as digital documents, images, voice recordings etc.).

Data curation involves organising and managing data from various sources. Data are collected and managed to respond to the needs of the respective user groups. Typically, they are organised into data sets and data catalogues. Data may also require **cleansing**, that is, correcting or removing records considered inaccurate or corrupt.

Data coding is a process of organising data into meaningful categories, based on a theory or specific assumptions. Both qualitative and quantitative data can be assigned numerical codes. Numeric coding is supported with various software (e.g. SPSS, Stata, SAS, R). Evaluators need to decide on the codes, usually in accordance with the evaluation purposes, questions and indicators. Labels are typically used to organise data and variables. Variables can be further grouped into sets that relate to the same phenomenon or feature (e.g. undernourished population living in a given region). When coding data, evaluators should be careful to preserve the original (raw) data and check against the various risks such as data disclosure and risks related to data quality.

Several policies or practices known as **data standards** help ensure the best possible data quality and interoperability. Respecting these standards can help maintain coding consistency and data use across different systems. **Data processing** means activities that generate information from data. Data can be processed manually or electronically, and through automation. It involves classification (organising data into meaningful categories), validation (checking that data are correct) and aggregation (combining raw data into cumulated pieces of data). The subsequent steps are related to data analysis, discussed in Section 3.4.

QUALITY ASSURANCE

The quality of evaluation data is always important. Ensuring data quality is especially challenging when data are collected on a large scale, from multiple

sources, and often from sources not designed with the intention to be used by an evaluation team – for example, big data. In these circumstances, attention needs to be given to data management.

Quality assurance processes make data useful for evaluation and decision-making purposes. This can be achieved through the thorough design of data architecture and work on data formats, curation, standards and coding, discussed below. Recognised attributes of high-quality data are as follows.

- **Validity.** Data have been cleansed of errors, making it correct and useful.
- **Completeness.** Missing data points are minimised or managed.
- **Consistency.** The same type of data stored in two different places match.
- **Accuracy.** Data are accurate – that is, the values are correct.
- **Uniformity.** As far as possible, different sources deliver data in the same format and values.
- **Integrity.** Data are kept consistent within their entire life cycle.

DATA VISUALISATION AND DASHBOARDS

An essential element of evaluation is **data visualisation**. There is no single approach to it, and data can be visualised either traditionally (e.g. by manual drawings) or in a digital format. Visualisations are also an important means of communication about an evaluation with its stakeholders, throughout the evaluation process. Visual representations are powerful tools to depict the complexity of the evaluand and can be key enablers of interactions between evaluators and the community. But there are also associated risks such as oversimplification or misleading visual metaphors. Visual data can also be manipulated to fit an intended perception of the audience. For instance, data visualisations can be manipulated with colours, scales, axes or cumulated graphs.

Visualisations can be as simple as a few lines on a paper sheet or as complex as advanced graphical representations of big data. The most common types are charts, tables, graphs, maps, infographics and dashboards.

Theory of change visualisations are important elements of evaluations, which use this tool to depict the complexity of interventions. A typical visualisation of a theory of change is focused on explaining the linkages between the evaluation objectives and their results – that is, outputs, outcomes and impacts – where feasible (see e.g. Figure 3.2.3 and Figure 3.2.4). These can be shown in linear or non-linear ways, depending on the complexity of an intervention. Apart from the standard MS Office (especially Word and Excel), there is a wide variety of software which could be helpful to depict the theory of change, such as:

- TOCO, a dedicated theory of change software;
- Theorymaker, which helps to develop a simple theory of change;
- Miradi, software particularly for conservation interventions;
- VUE, which supports the visual understanding environment.

Data dashboards are an increasingly popular means for navigating the evaluation data landscape, particularly when complex quantitative data are needed. They can be created in simple formats (e.g. with Excel) or include very advanced and real-time data applications, using a dedicated business intelligence software. An effective data dashboard enables a quick presentation of information that is most relevant for the evaluation and decision-making.

Several types of visualisations can be used for this purpose, reconfigured and combined, according to the evaluation's needs. **Infographics** differ from data dashboards in that they are static displays of the selected data to illustrate a specific topic of concern. They cannot be modified by the user once released. Some useful software includes:

- Tableau, popular software for interactive dashboards;
- Google Looker Studio, web-based data visualisation and dashboard tool;
- Microsoft Power BI, advanced business intelligence visualisation;
- Qlik, data visualisation software;

- Klipfolio, web-based data visualisation and dashboarding software;
- Shiny, R package for interactive web applications for data analysis;
- Dash Enterprise, Python framework for building analytical web applications;
- SAS, advanced statistics software with visualisation functions;
- ArcGIS, a pioneer of the visualisations with maps;
- QGIS, an open-source tool for visualisations with maps.

SEE: *Useful blogs and catalogues about data visualisation include* EvergreenData*;* Michelle Laurie Rants and Raves*; and* The Data Visualisation Catalogue.

SECURITY AND SAFE HANDLING

Intellectual property rights issues may arise with the use of data generated through the evaluation process. A common practice in the context of evaluation is the obligation of non-disclosure of data and other products related to evaluation, typically executed through a non-disclosure agreement between the evaluator and evaluation commissioners. This means that an evaluator, even if data collection and management involves significant personal efforts and often individual creativity, cannot make further use of the data beyond the evaluation contract. Exemptions are possible, if an agreement is reached between the evaluator and the commissioner in this respect.

SEE: Chapter 4 *for more guidance on legal and ethical issues connected with data collection and management.*

The EU promotes several **instruments dedicated to data standards**, such as the General Data Protection Regulation (GDPR, Regulation 2016/679), Open Data Standards and the INSPIRE Directive (Directive 2007/2/EC) which is dedicated particularly to spatial data.

NOTE: *The GDPR (*679/2016*) covers data protection; Directive* 680/2016 *covers data protection in the area of police and justice; Regulation (EU)* 2018/1725 *covers processing of personal data by EU institutions, bodies, offices and agencies.*

At the global level, ISO TC/69 is another useful approach that helps organise data for interoperability. By using these standards, an evaluator not only ensures the application of well-trusted data formats, but also allows for a more meaningful contribution to decision-making and agendas, which builds on many sources of data.

SEE: *ISO/TC 69* Applications of Statistical Methods*; EC* INSPIRE Knowledge Base*; and EC* data.europa.eu.

3.3.5 Data collection in contexts affected by fragility, conflict and violence

Various difficulties may arise when evaluation is in need of data generated in fragile contexts. These situations often limit the capacities of evaluators and stakeholders to access relevant data sources and respondents. Evaluators may not be able to travel to remote areas or those where their lives could be at risk. Novel ICT-driven technologies increasingly often enable data collection in fragile contexts (Hassnain, 2019; Hoogeveen and Pape, 2020). When face-to-face meetings are to be avoided, the following data collection methods can be used:

- mobile phones and tablets;
- geospatial data sources;
- online surveys;
- phone surveys;
- visual narratives (video, photography).

Further challenges of data collection in fragile contexts are associated with the lack of trust between stakeholders (including lack of trust in evaluators); instability, which could jeopardise an established data flow; and the limited or absent culture of learning, which can undermine progress in decision-making. Before starting an evaluation process, an evaluator should investigate the situation and assess the relevant risks.

Mitigation strategies could be based on triangulation, where similar types of data are obtained from multiple sources. A good backup plan may be necessary which involves advanced ICT solutions. Evaluators and enumerators may also need to have strong negotiation and conflict resolution skills, and an understanding of gender and conflict sensitivity which can be acquired through dedicated training and practice (Hassnain, Kelly and Somma, 2021). Apart from these, the use of existing data sets and benchmarking and proxies could be alternative solutions when data cannot be obtained directly.

SECTION 3.4

Data analysis

Data analysis is the process of interpreting and understanding the data collected during the evaluation; it is a critical activity that serves to transform raw data into findings, conclusions and recommendations. It enables the identification of patterns and trends, and allows predictions to be made. It can also be used to test hypotheses and evaluate results.

There are a variety of data analysis methods available, each with its own strengths and weaknesses and each best used with particular types of data and information (e.g. quantitative, qualitative, primary, secondary) and in particular evaluative designs (e.g. experimental, quasi-experimental and non-experimental). Some of the methods most commonly used in European Commission (EC) evaluations are discussed here.

This section looks first at the analysis of quantitative data, describing the procedure to be followed and the different types of statistical analysis. The next subsection describes methods for analysing qualitative data using software; this is followed by a discussion of new and innovative data science and data analytics tools that are coming into use. A discussion of mixed methods follows. The section concludes with a brief summary of the different sources of bias in data collection and types of errors – false positives and false negatives and confusion matrices.

3.4.1 Quantitative data: statistical analysis

Statistical analysis involves the application of mathematical and statistical techniques to understand and draw conclusions from quantitative data. Statistical analysis can be used to understand both primary and secondary data.

Secondary data are becoming more relevant for evaluations, given the increase in the volume and accessibility of data generated by research organisations, public administrations and non-profit groups (see Box 3.3.1). Having access to effective methods of analysing that data are important for evaluation teams that want to make use of the wealth of available information.

MAIN TYPES

Some of the most common types of statistical analysis are described below.

SEE: *Bevans (2020) for a thorough discussion of choosing the appropriate statistical tools.*

- **Descriptive statistics** are used to describe the data. This type of analysis can be used to calculate measures of central tendency (such as the mean or median) and measures of dispersion (such as the standard deviation or range). Descriptive statistics can be used to create graphs and charts. For example, in the context of an evaluation of an intervention aimed at reducing poverty, descriptive statistics can be used to calculate the mean and median incomes and examine the distribution of income levels of households before and after the intervention (i.e. establishing a baseline and an endline).
- **Inferential statistics** are used to make inferences from the data. This type of analysis can be used to test hypotheses and estimate population parameters. Inferential statistics are based on a few assumptions, and thus are not always accurate. For example, in the context of testing the hypothesis that an intervention had a significant impact on reducing poverty, inferential statistics could be used to estimate the size of the impact the intervention had on poverty reduction.
- **Regression analysis** is a type of inferential statistics used to identify relationships between variables and make predictions about future values. Regression analysis could be used to predict the amount of poverty reduction that would/should result from an increase in income by identifying the relationship between income and poverty reduction.
- **Univariate, bivariate and multivariate analyses** look at the kinds of relationships that can exist between different kinds of data.

SOURCE: *This material is drawn from unpublished material by Rick Davies (2023).*

- **Univariate analysis** focuses on one measure alone – for example, the nutritional status of children as measured by height for weight. The data on that measure can be analysed in terms of the central tendency (mean, mode, median), the maximum range of values or the shape of the distribution of those values.
- **Bivariate analysis** focuses on one-to-one relationships – for example, the relationship between nutritional status of children and family size. This kind of relationship can be summarised in a 2×2 cross-tabulation, a scatter plot or a simple linear correlation. In all cases, statistical tests can be applied to identify the significance of any relationship relative to a chance occurrence. Observations far from the average trend (outliers) can also be identified and discussed.
- **Multivariate analysis A** focuses on many-to-one relationships – for example, identifying which of the many different attributes of a household may be contributing to the nutritional status of the youngest child. These relationships can be analysed using statistical methods such as multiple regression, machine learning algorithms, or configurational analysis methods such as qualitative comparative analysis.
- **Multivariate analysis B** focuses on one-to-many relationships – for example, identifying which of the many different changes in a household's well-being might have been caused by joining a savings and credit group. The same types of analytical tools can be used as with many-to-one relationships (multivariate

analysis A). Analysis of the effects of a cause is less common than analysis of the causes of an effect; it is relevant in interventions where a diversity of effects might be expected because of the nature of the intervention or because of the diversity of people and contexts in which that intervention is taking place.

- **Multivariate analysis C** focuses on many-to-many relationships – for example, the relationships between different programmes in a portfolio, between different interventions within a single programme or between different households within a community. These kinds of relationships can be analysed using statistically based cluster analysis; ethnographically based pile or card-sorting exercises undertaken by one or multiple participants (an ethnographic approach); network analysis, using specialised social network analysis software; and case study methods.

- **Significance tests** are used to test hypotheses about relationships between variables. They are commonly used to determine whether the results of an analysis are due to chance or whether they are statistically significant (i.e. how much the results of the analysis are valid). This type of analysis can be used to compare the results of a study to a control group, the results of different studies or the results of a study to a theoretical model. For example, in the context of an evaluation of an intervention aimed at reducing poverty, significance tests could be used to compare the mean income of households before and after the intervention. Significance tests could also be used to compare the proportion of households below the poverty line before and after the intervention. In a quasi-experimental setting (see discussion under Subsection 3.2.2), a significance test could be used to compare the poverty rates of households that benefited from the intervention to the poverty rates of households that did not.
- **Analysis of variance** is a significance test that can be used to compare the results of an intervention across different subpopulations. For example, if an intervention is being evaluated for its impact on poverty, analysis of variance could be used to compare the results of the intervention by age group, sex, geographic location etc.
- **Time series analyses** are used to examine data over time. This type of analysis can be used to identify trends, to make predictions about future values and to identify relationships between variables. For example, in the context of an evaluation of an intervention aimed at reducing poverty, time series analysis could be used to examine poverty rates over time. Time series analysis could also be used to predict how poverty rates would change in the future if the intervention were continued.

STRENGTHS AND WEAKNESSES

Statistical analysis is a powerful tool that can be used to understand data. However, it is only as good as the data on which it is based. In order to make accurate inferences, it is important to **use relevant, high-quality data**. Some particular data limitations to be aware of follow.

Compared to primary data, **secondary data** may be more economical and time saving, but may not always include the specific information an evaluation needs. For example, a database of employment statistics may not contain information on the specific skills workers have. This lack can make it difficult to assess whether an intervention has had a positive impact on the skills of the workforce.

Furthermore, the **available data may be too broad** to be relevant to a specific local context of the evaluation. For example, national data on poverty rates may not be relevant to a specific evaluation of an intervention in a rural area. This is because the data may be aggregated at too high a level and may not take into account the specific circumstances of the rural area.

There may be other limitations related to **lack of transparency on data collection** procedures whereby, for example, data that are collected by some agencies may be subject to political biases. Additionally, secondary data may **not be available in a timely manner**, which can make it difficult to use for evaluations.

3.4.2 Software-assisted qualitative data analysis

Qualitative data analysis software is being increasingly used in evaluations. It enables evaluators to organise, analyse, and interpret data from interviews, focus groups, and other qualitative sources. Using these software makes qualitative data analysis more structured and its conclusions more transparent and evidence-based. Some of the main features of qualitative data analysis software include the following.

- **Helping code data from interviews and focus groups.** Codes are typically assigned to excerpts to categorise and organise them. Codes can be descriptive (e.g. age, gender, location) or analytical (e.g. themes, patterns). Coding typically involves reading through transcripts of interviews or focus groups and assigning codes to excerpts that are relevant to evaluation questions.
- **Identifying patterns and themes.** For coded data, these software support generating lists of codes, identifying patterns and relationships between codes, and creating summary tables and charts.
- **Tracking changes over time/location.** This can involve creating timelines to track progress, documenting changes in codes and coding schemes over time and/or location, and incorporating qualitative data into quantitative analysis.
- **Comparing data from different sources.** This can involve creating codebooks to compare data from different sources, aligning codes across sources and creating comparative reports.
- **Generating reports.** This can involve creating text summaries, tables and charts to communicate findings.

QUALITATIVE DATA ANALYSIS TOOLS

Many qualitative data analysis software exists, and each has its strengths and weaknesses. Below is a list of the more popular software; their features are summarised in Table 3.4.1.

- Taguette offers an easy-to-use software with a wide variety of import and export options. It is also free and open source. However, it does not have any visualisation capabilities.
- ATLAS.ti enables importing of survey data, as well as data visualisation tools. It also supports mixed-methods data analysis. However, the interface is complicated, and classification of data can be challenging.
- MAXQDA is used for analysing data from different sources (including Twitter) and analysing mixed-methods data.
- NVivo is easy to learn but has a complicated interface and does not have an auto code option.
- Dedoose is used for analysing text, audio and video files. It is web-based and supports teamwork in real time. However, it has a complicated web-based interface with many options for data analysis which may be overwhelming for some users. Data

TABLE 3.4.1 Summary of qualitative data analysis tools

Software	Text import	Survey import	Audiovisual import	Mixed methods	Data visualisation	Inter-rater reliability	Auto coding	Team work
Taguette								
ATLAS.ti								
MAXQDA								
NVivo								
Dedoose								
QDAMiner								
Qcoder								

are stored on the cloud which may not be optimal in all cases.

- QDA Miner is a good choice when support for a wide variety of audio and text formats is needed. The software supports mixed-methods data analysis. However, it does not have a strong community of users, even though online automated training tutorials are available.
- Qcoder would be a good choice for those who want to use R for their data analysis. However, it requires knowledge of R to use.

STRENGTHS AND WEAKNESSES

Using qualitative data analysis software in evaluations is helpful in several ways. It leads to a more structured and transparent analysis. It is not, however, a substitute for actual data analysis.

Qualitative data analysis is an iterative and inductive process that requires human judgement and interpretation. The use of software can help automate some aspects of the data analysis process, but it is ultimately up to the evaluator to make sense of the data and draw conclusions.

Additionally, preparing and entering data in software can be time-consuming. Depending on the size and complexity of the data set, it may be necessary to invest significant time to get the data ready for analysis. This is an important consideration when planning an evaluation, as it can affect the overall timeline and budget.

3.4.3 New data science and data analytics tools

There is an increasing acknowledgement of the potential for data science and data analytics tools to play a role in evaluations. While there has been some reluctance to embrace these tools in the past, there is a growing recognition of their potential to provide insights that would otherwise be unavailable. One of the key advantages of data science and data analytics tools is their ability to extract meaning from data that are unstructured or difficult to access. This is particularly relevant in the evaluation of interventions that rely heavily on qualitative data, such as media articles or progress reports.

Another key advantage of data science and data analytics tools is their ability to detect patterns and trends that would be difficult to identify using traditional methods. This is important, as development interventions are designed to bring about long-term changes that may be difficult to measure.

There are some challenges that need to be considered when using data science and data analytics tools in the evaluation of international development interventions.

- Some data science and data analytics tools **require access to large amounts of data**. This can be a challenge in many developing countries, where data are often scattered and of poor quality.
- Data science and data analytics tools **can be complex and require specialised skills** to use effectively. This can be a challenge in many evaluation contexts, where resources are often limited.
- Data science and data analytics tools **can be misused or misinterpreted**. This is a challenge in any evaluation context, but it is particularly relevant for machine learning tools as their complexity sometimes prevents proper interpretation of the validity of their findings.

The term 'machine learning' covers a broad category of data analysis with a common type of process. They all involve the use of automated algorithms (i.e. documented procedures) to incrementally search for and find the best available solution to a problem. The problem may be to find the best combination of variables (e.g. measures of economic performance) or categories (e.g. types of companies) to accurately predict a performance measure of one kind or another (e.g. company market share performance). Alternatively, the problem may be to find the best way of grouping a set of objects or events, such that the member of each group has more commonalities with others in the group, compared to those in other groups – for example, types of small enterprises.

Machine learning algorithms can use many different types of inputs: numbers, words, images or even sounds. They are now widely used by all types of

companies, as well as research institutions and government services. Evaluation teams should be expected to have at least some basic knowledge of machine learning.

For a given machine learning task, such as developing a good predictive model of what kinds of interventions might best lead to a particular desired outcome, there will be multiple choices available as to the type of algorithm that could be used. In addition, the performance of the selected machine learning algorithm will depend on the choice of settings (parameters) governing how the algorithm will work.

If evaluation teams are using machine learning algorithms, they should be able to explain what type of algorithm was chosen and why, and what parameters were used and why.

With some evaluations, public understanding and trust in the findings will be particularly important. In these situations, the evaluation team should be able to put forward a simple explanation of how a particular algorithm works. They should also be able to respond to challenges that might be made about the biased nature of the data that the algorithm used and its effects.

SEE: *Kotu and Deshpande (2014); O'Neil (2016).*

MAIN TYPES

Some of the main types of data science and data analytics tools that can be used in this context include the following.

- **Natural language processing tools[1].** These tools can be used to automatically extract information from text-based data sources, such as media articles or progress reports.
  - **Text mining tools.** These tools can be used to identify relevant documents and sources of information from a large corpus of data. They can be used to automatically identify and categorise documents, and to identify patterns and trends in the data.
  - **Social network analysis tools.** These tools can be used to analyse data from social media platforms, such as Twitter or Facebook. They can be used to identify patterns and trends in how people are talking about a particular intervention.
- **Imagery analysis (satellite).** Imagery analysis tools can be used to automatically extract information from satellite images. For example, they can be used to identify patterns and trends in land use, land cover and land management practices.
- **Other general machine learning tools.** These tools can be used to extract information from data sources that are difficult to access or are unstructured. They can be used to identify patterns and trends in data and are the main types of machine learning algorithms:
  - **Regression.** This type of algorithm is used to predict continuous values, such as a future stock price or the likelihood of someone developing a disease. Linear regression is the most popular regression algorithm, but there are also more sophisticated methods, such as support vector machines.
  - **Classification.** This type of algorithm is used to predict which category a particular instance belongs to, such as whether an email is spam or not. There are many different classification algorithms, but some of the most popular are decision trees, k-nearest neighbours, and Naive Bayes.
  - **Clustering.** This type of algorithm is used to group similar instances together. For example, a clustering algorithm could be used to group customers together based on their purchasing habits. K-means clustering is the most popular clustering algorithm, but there are also other methods, such as hierarchical clustering.

APPLIED EXAMPLES

The range of application of these tools is wide and largely unexplored. The following indicates how some can be used in a variety of intervention contexts.

[1] There have been important recent developments in this area which are too large scale in their implications (even for data analysis alone) to be addressed here.

- **Electricity access intervention.** Use satellite imagery to look at the change in the number of lights visible at night (which could alternatively be used as a proxy for a change in increased economic activities, safety, health facilities etc.). For example, night-time light data were used in a UK-funded intervention implemented by the United Nations Office for Project Services in Sierra Leone.
- **Technical and vocational education and training:**
  - Use phone records data to determine the approximate location and movements of people who have completed a vocational training programme. These data can then be used to assess the impact of the programme on mobility and migration.
  - Use electronic transaction data to estimate the number of people who have found formal jobs after completing the training programme.
- **Sector evaluation.** Use text analytics to aggregate indicators such as the number of people reached and the number of people benefiting from large bodies of unstructured text (progress reports etc.).
- **Poverty alleviation intervention.** Use text analytics to look at changes in the content of poverty-related news articles over time.
- **Food security intervention:**
  - Use satellite imagery to look at the amount of green vegetation in an area, which is a proxy for food production.
  - Use phone data to estimate the impact of a food security intervention by looking at changes in the movement patterns of people before and after the intervention.
- **Environmental conservation intervention.** Use satellite imagery to estimate the Green Vegetation Index in an area, which is a proxy for the health of the environment.
- **Infrastructure development budget support intervention.** Use satellite imagery to extract the number of kilometres of new roads.

3.4.4 Using mixed methods

Mixed-methods evaluations are those that use both quantitative and qualitative data to paint a more complete picture of what is being evaluated. Several kinds of qualitative data analysis software offer features that allow the integration of mixed methods.

Mixed-methods evaluations enable the triangulation of data. This means that if there are discrepancies between the quantitative and qualitative data, they can be investigated and resolved. This can lead to a more accurate understanding of the situation.

Mixed-methods evaluations can also be useful in situations where there is a need to understand both the process and the outcomes of an intervention. Quantitative data can be used to measure outcomes, while qualitative data can be used to understand the processes that led to those outcomes.

By using a mixed-methods approach, evaluators can make use of the strengths of both quantitative and qualitative data. This can help to ensure that the findings of an evaluation are accurate and reliable, and that the data collected are relevant and useful.

The main benefits of mixed methods are as follows:

- Mixed methods allow for triangulation of evaluation findings (if there is convergence, there is greater validity; if there is incoherence, there is a need for analysing reasons).
- Diverging results call for reconciliation through further analysis.
- Results from one method help develop the tools/sample/instrumentation of another – that is, there are feedback loops between the two data sets.
- The two data sets are complementary, leading to broader, deeper understanding.
- Mixed methods incorporate a wider diversity of values using different methods.

USE QUANTITATIVE DATA TO SHAPE QUALITATIVE DATA COLLECTION

One way to use mixed methods is to use quantitative data to shape qualitative data collection. An

evaluation might use a survey to collect quantitative data on beneficiaries' perspectives and then use interviews and focus groups to collect qualitative data to explore those opinions in more depth. This can be a very effective way to collect data, as it allows the evaluation to focus qualitative data collection on the areas that are most important based on the quantitative data. It can also help to identify areas where further research is needed.

For example, if a survey finds that a certain group of people is more likely to experience a certain type of problem, qualitative data can be collected from this group to explore the issue in more depth. This mixed-methods approach can be used to collect data from hard-to-reach groups or to explore sensitive topics. Similarly, qualitative data can be used to shape quantitative data collection. For example, findings emerging in interviews could lead to the elaboration of a survey or more complex designs with combinations of QUAL-QUAN-QUAN data collection employed.

USE QUALITATIVE DATA TO TRIANGULATE QUANTITATIVE DATA

Another way that mixed methods can be used in evaluations is by using qualitative data to triangulate quantitative data. This means that qualitative data are used to check and confirm the findings of the quantitative data. This can be done, for example, by conducting interviews with people who have been surveyed. This mixed-methods approach can help ensure that the findings of an evaluation are accurate and reliable, especially in cases where the quantitative data collection relies on non-statistical sampling methods.

3.4.5 Sources of bias in data analysis

The analysis of qualitative and quantitative data can easily be skewed by different kinds of bias. Very generally speaking, a bias is 'an unjustifiable tendency to lean in a certain direction, either in favour of or against a particular thing'. In addition to the potential biases in data collection described in Box 3.3.2, evaluators can sometimes be biased in the way they analyse data, and the readers of evaluations may be biased in the way they interpret evaluation findings. Both evaluators and commissioners of evaluations need to be aware of the kinds of bias that may be present. Some of the more widely recognised forms of bias are described below.

- **Confirmation bias** occurs when the person performing the data analysis wants to prove a predetermined assumption. He/she will keep analysing the data in different ways until this assumption can be proven – for example, by intentionally excluding variables, or particular observations, from an analysis.
- **Apophenia** is the tendency to perceive meaningful patterns within random data.
- **Anchoring bias** is the tendency to rely too heavily on the first piece of information offered when making decisions.
- **Halo effect** is when an impression is formed due to a single characteristic which then influences multiple judgements or ratings of other unrelated factors.

3.4.6 Sources of errors in data analysis: the confusion matrix

If evaluators find a significant correlation or association between one event and another, for example a particular intervention and a particular outcome, they should also be able to identify any limitations with the finding and how it might affect conclusions that are made based on that finding. Most correlations/associations are not perfect; there will be exceptions, and the scale and nature of these exceptions need to be identified and reported on by an evaluation team. Errors can have consequences, including inequities of outcomes.

A confusion matrix is a tool that can be used for this purpose, and more. Both commissioners of evaluations and evaluation teams should be familiar with the confusion matrix. An abstract example is shown in Figure 3.4.1.

FIGURE 3.4.1 An empty confusion matrix

	Observations	
Prediction	Intervention outcome is **present**	Intervention outcome is **absent**
Intervention X is **present**	True positive	False positive / Type 1 error
Intervention X is **absent**	False negative / Type 2 error	True negative

When a confusion matrix is used, each of the four grey cells will have a value representing the number of cases or observations that fit the cell description. For example, the true positive cell will show the number of cases (households, businesses, villagers, interviews etc.) where intervention X was present, and the intervention outcome was also present. Similarly, the false positive cell will show the number of cases where intervention X was present, but the interventions outcome was absent.

There are a range of performance measures for assessing how well two events such as X and Y are related to each other, given the numbers in the cells. Different circumstances often need different performance measures.

SEE: Simple guide to confusion matrix terminology *(Data School website, 2014).*

FALSE POSITIVES AND FALSE NEGATIVES

The two types of error represented in this matrix are widely recognised:

- A false positive, also known as a Type I error, is present when the outcome is not present even though it is predicted to be present. For example, imagine a test for COVID-19 which says that a person has COVID-19 but further and more detailed diagnoses indicate otherwise.
- A false negative, also known as a Type II error, is present when the outcome is present even though it has not been predicted. For example, a test for COVID-19 says that the person does not have COVID-19, but subsequent more detailed diagnoses indicate that the person does.

Different types of errors can be acceptable in different types of situations – a point that evaluation teams should be aware of and may need to highlight. In the above example, the Type 2 error could have much more dramatic consequences than the Type 1 error; but there could be circumstances where Type II errors may have more dramatic effects comparatively – for example, a positive mammogram screening test result for breast cancer that erroneously led to a radical mastectomy. When an important relationship is identified by an evaluation team, they should be able to also identify the scale and consequences of both types of errors in that context.

POSITIVE DEVIANTS

Not all false positives are bad. An evaluation team might quite reasonably develop a predictive model of the conditions under which the outcome **does not occur**. It could be quite a good model covering a large proportion of the cases that perform poorly, and the model could also have a high consistency. But the few cases which are false positives could be potentially important and warrant further investigation. For example, as was found in Vietnam in the 1990s in a community where poverty was widespread and so was childhood malnutrition, a few poor households with above-average childhood nutrition status could be a source of potential useful feeding practices for other households in the community.

SEE: *Marsh et al. (2004).*

COVERAGE AND CONSISTENCY

Coverage and consistency are two technical terms borrowed from qualitative comparative analysis (discussed in Subsection 3.2.4), but they have wider applicability (and other equivalent terms). How these two measures are related is shown in the confusion matrix in Figure 3.4.2.

FIGURE 3.4.2 A populated confusion matrix

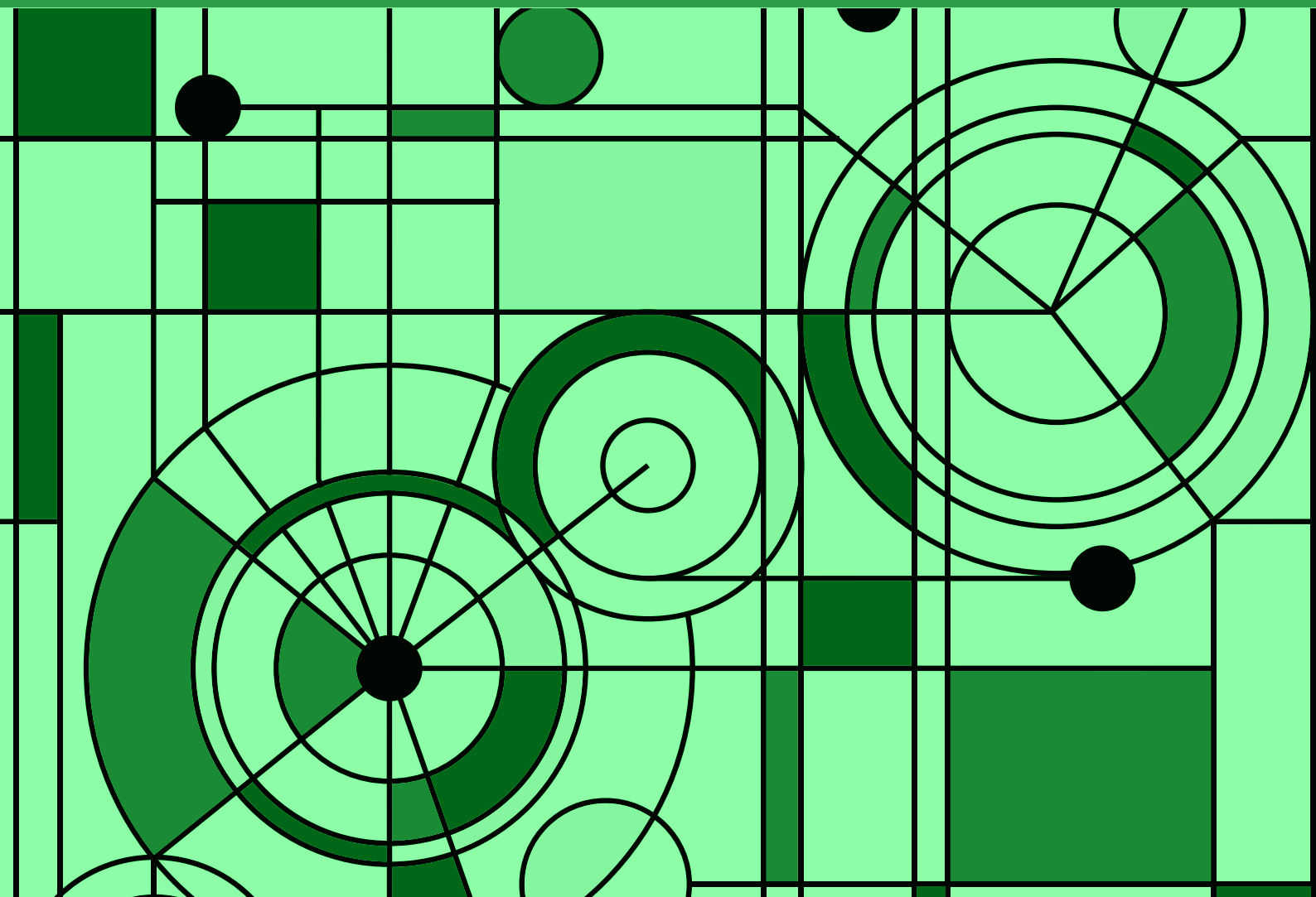
		Desired outcome is...		
		Present	Absent	Consistency / precision / positive predictive value
Intervention is	Present	N= 20	N =5	= (20/(20+5)) = 80%
	Absent	N = 35	N= 45	
	Coverage/recall/sensitivity	= (20/ (20+35)) = 57%		

An evaluation team might find that with a particular intervention the outcome is present most of the times when the intervention is present. The measure that describes this relationship is consistency and is calculated as 80 per cent in the example in Figure 3.4.2. But when the evaluation team looks at all the times the outcome is present, many of these cases are occurring when the intervention is not present. The measure that describes this relationship is coverage and is calculated as 57 per cent in the example in Figure 3.4.2.

It is important to note that the significance of these measures will vary according to context. For people investing in the stock market, they may want a predictive model that has a wide coverage, but they may not be too worried if it does not have a very high consistency – just so long as the predictions are more often right than wrong. On the other hand, a doctor planning an intervention with a critically ill patient will be very concerned about the likelihood that the prediction about the intervention is as near to 100 per cent correct as possible, and not be too concerned that the model only applies to a quite narrowly defined set of cases – that is, has low coverage.

4

Ethics in evaluation



What is this chapter about?

This chapter underscores the fundamental nature of ethics in evaluation, details ethical principles and standards, and delineates ethical considerations across and throughout the evaluation process.

How will this help you in your work?

This chapter suggests practical ways to address ethical considerations and the best possible mitigation actions in your evaluation practices. It presents guidance for handling some of the most common ethical issues when consulting and interviewing people especially in contexts of fragility, crisis, conflict and violence.

For definitions of key terms used in this handbook, refer to the glossary.

Worldwide, a common set of fundamental ethical principles underpins all evaluations; every evaluation society and organisation ascribes to these, albeit with slight variations. These principles are embodied in the European Union's (EU's) Evaluation Policy, which states that:

> The rights and dignity of all evaluation stakeholders are respected. The design of an evaluation must consider and address potential ethical challenges which may arise. Evaluations, and the evaluators, should respect the rights and dignity of respondents, programme participants, beneficiaries, and other evaluation stakeholders. They must explain and preserve confidentiality and anonymity of participants, where sought or provided. Those who partake in an evaluation should be free from external pressure, and their involvement should not disadvantage them in any way (EEAS and EC, 2014, p. 14).

4.1 EU ethical principles

Figure 4.1 illustrates and Table 4.1 synthesises **six fundamental ethical principles** underpinning the EU Evaluation Policy and explains what each means in practice.

That, above all, is the key takeaway about ethics in evaluation: it is not just idealistic premises or promises, but **practical, pragmatic, real-world actions** taken to ensure that all stakeholders

FIGURE 4.1 Ethical principles

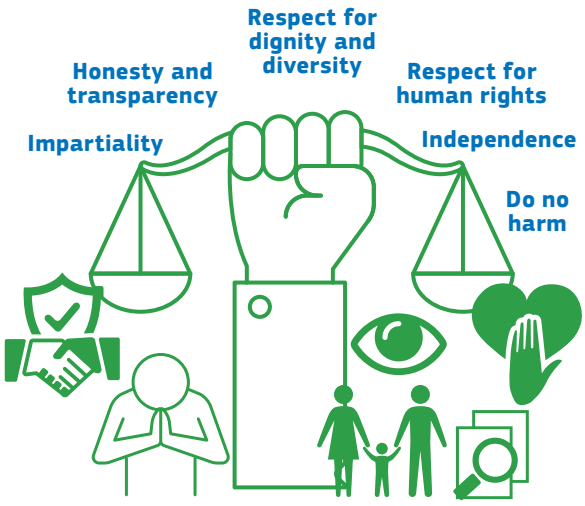


TABLE 4.1 Fundamental ethical principles

Principle	Explanation	Good practice example
Honesty and transparency of the entire evaluation process	Evaluation purpose, procedures, data, findings and limitations are transparently and accurately represented and communicated to stakeholders; and evaluators justify their judgements, findings and conclusions.	• Sources are cited, and biases and limitations are acknowledged and explained, in evaluation reports. • Evaluators do not bias the evaluation to receive further commissions from the client, nor do they accept gifts or payments intended to influence evaluative judgement.
Respect for human rights when engaging with individuals or groups in the evaluation	Evaluators respect and protect the rights and welfare of individuals and communities in accordance with the United Nations Universal Declaration of Human Rights and the European Convention on Human Rights.	• Evaluators comply with legal and safeguarding codes when conducting interviews, especially when interviewing children, younger and older people, women, survivors of gender-based and other kinds of violence etc. • Evaluators ensure that prospective participants and interviewees are free to choose whether to participate in the evaluation and are aware of their rights in this respect. This is also a legal obligation for the Commission under Article 39 of the General Data Protection Regulation.
Respect for dignity and diversity of individuals and societies	Evaluators acknowledge the value of individuals, respecting differences in culture, customs, religious beliefs and practices, gender roles, sexual orientation, disability, age and ethnicity.	Evaluators strive to be culturally and contextually knowledgeable and sensitive, designing evaluation instruments appropriate to the values of individuals and societies (e.g. culture, heritage) and to the context, especially in fragile and conflict-affected environments.
Do no harm: people must not be exposed to harm as a consequence of the evaluation	Evaluators consider the broader context of the intervention and its underlying risk factors and take action to minimise or mitigate the risks of potential harm or negative effects on individuals, the economy and the environment.	• Evaluators respect the anonymity of key informants when this is requested to protect their safety. • Evaluators adapt their plans to address concerns identified by the most recent conflict analysis. • Evaluators identify any potential conflict triggers before gathering people for a group discussion. • Evaluation managers consider any possible negative effects in a given context before approving field visit plans, and take action to address these in consultation with the evaluation team.
Impartiality: the evaluation provides a comprehensive and balanced analysis	Evaluators take account of the views and experiences of a diverse cross-section of stakeholders to avoid or minimise bias and to enhance the credibility of the evaluation.	• Policymakers and intervention managers should not evaluate their own work. • If evaluators have been unable to reach certain stakeholders (e.g. inaccessible population groups), this is clearly stated as a methodological constraint, and the findings qualified accordingly.
Independence: evaluators must be free of bias and conflicts of interest	Evaluators feel free from any external pressure in producing meaningful evidence in support of institutional learning and effective and accountable decision-making.	Evaluation commissioners and managers must not impose restrictions on the scope, content, comments and recommendations of the evaluation process and outputs.

in evaluation – evaluators, commissioners of evaluations, donors and other partners, beneficiaries, users – act and are treated fairly.

This chapter presents ethical considerations and mitigating actions for various evaluation participants, activities and contexts. And because ethics should guide the entire evaluation process and be explicit in all quality assurance procedures and reviews, Table 4.2 summarises ethical considerations to take into account throughout the different phases of an evaluation.

NOTE: *For resources on ethical guidance, see the International Development Evaluation Association's Code of Ethics for evaluators and commissioners, and for evaluation as a profession (IDEAS, 2013); and the United Nations Evaluation Group's Ethical Guidelines for Evaluation which aims to ensure that an ethical lens informs day-to-day evaluation practice (UNEG, 2020).*

TABLE 4.2 Ethical considerations throughout the evaluation process

Consideration	Possible mitigating action / good practice
Preparatory phase	
• Is an ethical (credible, impartial, unbiased) evaluation feasible? • Are the selected evaluation team members, including the field enumerators, free from any conflict of interest with regard to any possible future developments and opportunities?	• Conduct an evaluability assessment (see Subsection 3.1.1) to determine if an ethical evaluation is feasible to capture learning or if an alternative review process should be considered (e.g. an after action review). • Ensure all evaluation team members have the requisite skills and their private interests do not conflict with the objectives, process, approach and outcomes of the evaluation exercise at present or in future. When recruiting, consider attributes such as socioeconomic status, sex, ethnic origin, age, values, religious beliefs, marital status and – in some contexts – identity politics. • Ensure effective procedures to resolve conflict-of-interest situations are available.
Inception phase	
• Will the proposed evaluation methods and tools meet ethical standards? • If working in contexts affected by fragility, conflict and violence, are the evaluation methods and tools conflict sensitive? Do they acknowledge and respect cultural and gender norms in the specific context? • Are evaluation team members fully aware of EC ethical guidelines and professional standards? • Does the evaluation team have a dedicated person for quality assurance? • Is there an ethical focal point and are the evaluators clear about whom to contact to discuss any ethical issues? • Is there an explicit assessment of ethical risks and proposed mitigation actions in the inception report? • Have clear and appropriate procedures and safeguarding measures been put in place to ensure informed consent within each method/tool, and have safeguards been considered in the evaluation design? • If applicable, are protection protocols outlined in the inception report for both vulnerable populations and evaluation personnel? • Does the inception report outline clear protocols for the storage and destruction of data after the evaluation?	• Design an evaluation framework based on a robust context analysis and taking key risks and assumptions, gender norms, values and beliefs, and conflict triggers into consideration. • Run safeguarding checks on interviewers, and plan for appropriate training of people carrying out data collection, professional exchange and ongoing supervision to ensure they have the necessary skills and knowledge of relevant codes of conduct. • Ensure that the rights and dignity of evaluation participants, including those managing and conducting evaluations, are acknowledged and respected as per international human rights conventions. • Ensure that the identity and confidentiality of evaluation participants and other stakeholders are protected throughout the evaluation process.

(continued)

TABLE 4.2 Ethical considerations throughout the evaluation process (continued)

Consideration	Possible mitigating action / good practice
Interim phase	
Are ethical considerations guiding day-to-day implementation and adaptation of evaluation methods?	• Design evaluation instruments appropriate to the values of individuals and societies, and to the context, especially in fragile and conflict-affected environments. • Respect accepted norms and behaviour when gathering data from vulnerable groups, such as women and children, especially in culturally sensitive areas. • For focus group discussions, ensure that groups with grievances against each other do not participate in the same discussion; instead, conduct separate interviews/discussions. • For interviews and surveys, ensure that sensitive questions are not asked, especially to an already traumatised community.
Synthesis phase	
• Is the final report credible from an ethical perspective? • Is there an explicit mention in the final report of ethical risks (including those reported in the inception report) and actual mitigation actions taken? • Does the report reflect a thorough contextual understanding throughout the design, implementation and reporting of results and recommendations? • Does the report demonstrate an understanding of power relations and inequality where appropriate? • Are the voices of the most vulnerable included? • Does the report discuss any potential negative effects of the findings, conclusions and recommendations?	• Ensure methodological constraints and mitigation measures taken are clearly stated in the methodology section of the final report. • Qualify any data omissions (e.g. of a particular part of the country or stakeholder group) in presenting relevant findings. • Ensure any intended and unintended positive and negative results of the intervention and their consequences are considered and reported in the final report.
Dissemination phase	
• Has ethical consideration been given to evaluation dissemination activities? • Are there actions for closing the evaluation learning loop with the relevant stakeholders, including in the communities where data were gathered?	Review which findings could be considered particularly sensitive and likely to put population groups at risk, consulting with staff in-country as necessary; consider putting those findings into a confidential document for the organisation's internal use, while releasing the rest of the report into the public domain.
Follow-up phase	
(For the evaluation manager:) Are there plans in place for follow-up on evaluation findings and recommendations?	(For the evaluation manager:) Ensure there is a follow-up plan for the evaluation findings and recommendations (see Section 2.7).
Quality assurance	
• Have the evaluation process and outputs been subjected to quality assurance? • Are the evaluation outputs inclusive and produced with high integrity, and without any bias or pressure?	• Ensure that quality assurance processes are in place. • Ensure that the final report fully represents the findings and conclusions of the evaluators and has not been amended without their consent.

NOTE: Considerations apply to all evaluation stakeholders, especially the evaluation manager and evaluation team.

4.2 Ethical standards and actions for evaluators

Asymmetrical power relations, the prevalence of donor-recipient modalities of thinking and acting, and cross-cultural differences make evaluation of international development strategies, policies, instruments, modalities or interventions difficult and subject to intricate ethical choices.

These choices are easier to assess when they are based on clear ethical standards derived from the principles set out in Table 4.1. Such ethical standards provide a **moral compass** – both for those who commission evaluations and for those who carry them out – in three critical ways:

- **They protect the rights and welfare of the individuals, groups or organisations that take part in or are consulted by the evaluation.** This means that when designing and carrying out the evaluation, the evaluation commissioners, managers and evaluators respect human rights and the differences in culture, customs, religious beliefs and practices of all stakeholders, and are mindful of gender roles, ethnicity, ability, age, sexual orientation, language and other differences.
- **They clarify the guiding principles and related conduct of those who commission and carry out the evaluation to ensure a credible and independent evaluation output.** For example, evaluation managers should establish measures for identifying and mitigating conflicts of interest for those who are to carry out evaluation, with consideration of possible future developments and opportunities.
- **They define the acceptable conduct and behaviour of those carrying out the evaluation.** For example, evaluation managers should determine whether evaluation teams understand the guiding ethical principles outlined in Table 4.1 and have the required qualifications, expertise and experience to conduct the evaluation sensitively.

> **NOTE:** *Any breach in professional conduct can undermine the integrity of the evaluation being carried out – and indeed undermine the overall evaluation function.*

Setting and following strict ethical standards ensures that evaluation is fully grounded in the specific context.

> **SEE:** Table 4.2 *for some ideas on how ethical standards could be established and maintained throughout an evaluation exercise.*

If followed appropriately, these standards improve evaluation quality and ultimately strengthen the contribution of evaluation by promoting equity and transformative change – such as shifting power dynamics; reducing exclusion and discrimination; and increasing the autonomy and participation of the most marginalised or those excluded on the basis of race, ethnicity or gender.

> **SEE:** *The Australasian Evaluation Society's Guidelines for the Ethical Conduct of Evaluations, which it calls a 'framework for discussing ethical issues, and for helping people to recognise and resolve particular ethical issues that arise in the course of an evaluation' (AES, 2013, p. 2).*

4.3 Ethics in engaging and protecting

> **NOTE:** *This section draws on ALNAP (2018). Also see OECD DAC (2022).*

The ethical starting point for all evaluations is to consider the different ways in which engagement in the evaluation process might affect those who take part in or are consulted by the evaluation. These evaluation stakeholders include the **individuals and communities contacted and consulted by evaluators** and who are subject to the power dynamics inherent in the evaluation process. Local citizens may feel pressured to engage with an evaluation process in their community regardless of their personal opinion about it and whether their participation might have negative repercussions or carry other risks.

Another set of evaluation stakeholders is the **local researchers and enumerators participating in the evaluation**, who are often exposed to greater risks than international evaluators, particularly if they

are engaged in fieldwork in insecure or hard-to-reach areas.

> **NOTE:** *According to the Aid Worker Security Database, a large majority of aid workers experiencing attack or violence are national staff. This figure remained high during the COVID-19 pandemic (Humanitarian Outcomes, 2021).*

Where development cooperation is targeting the poorest, by definition they are also likely to be some of the most **vulnerable members of society**. Evaluators should take account of the following:

- **Beware of publicly exposing the views of individuals and groups most at risk**, making them more vulnerable to reprisals and abuse by powerful actors. Where this is a particular concern, consider how that information could remain confidential, while still informing the overall evaluation conclusions and recommendations.
- In engaging with survivors of sexual violence, take account of the World Health Organization recommendation that **basic care and support for survivors of sexual violence must be available locally** before any activity is carried out that is likely to involve an individual disclosing their experience of sexual violence – for example, in an interview or focus group discussion (ALNAP, 2018). If it is not possible to fulfil this requirement, the evaluators should not knowingly consult with survivors of sexual violence.

> **NOTE:** *This consideration is relevant to current discussions around how to address sexual violence in armed conflicts.*

- **Design evaluation methods so they are sensitive and appropriate to the issues to be explored**, for example, some protection issues related to individuals may be better explored in individual interviews rather than focus group discussions to avoid triggering feelings of shame, stigma or fears of recrimination.
- When recruiting interviewers and data collectors to work with vulnerable people – particularly in contexts where there are high levels of suspicion and distrust – **consider personal attributes that are likely to invoke trust** and facilitate relationship building with prospective interviewees (e.g. socioeconomic status; religious beliefs, which are extremely important in some contexts; and marital status). This may also be an important consideration for the safety of the interviewer/data collector. In contexts of fragility, conflict and violence, or gender-based violence, **attention should also be paid to identity politics**. Besides ensuring 'do no harm', being aware of the interviewers' and data collectors' identity can affect their motivation and attitudes towards their work.

> **EXAMPLE:** *An unmarried young woman training an older married women of a different ethnic group to administer a survey about maternal health may not be taken seriously or may be perceived as intimidating.*

Above all, **ensure that all those who carry out interviews are sufficiently well-trained to**:

- **understand and identify protection risks**, such as how participation in an interview could put an interviewee at risk, and know how to respond – for example, not to carry out the interview in such a case;
- **interview sensitively**, for example, avoid asking questions that will encourage the interviewee to relive a traumatic experience – such as in the case of a village having been attacked, avoid asking villagers exactly what happened in the attack, but instead focus on their needs after the attack and whether they received timely and relevant assistance and support;
- **respond appropriately should a protection issue arise during the interview** – for example, recognise when and how to stop an interview if the informant becomes distressed, or if a safe space is compromised by the arrival of an individual the interviewee does not trust or does not know;
- **ensure that interviewees are aware of and understand the available feedback mechanisms**, should it emerge that misconduct is to be reported;
- **recognise when the risks of carrying out the evaluation**, in terms of the safety of the interviewer(s), mean that the fieldwork should be aborted.

The evaluation team has a responsibility towards both the European Commission (EC) and the groups and individuals involved in or affected by the evaluation:

- Interviewers must ensure that they are familiar with and respectful of the beliefs, manners and customs of interviewees.
- Interviewers must respect people's right to provide information in confidence and ensure that sensitive data cannot be traced to their source.
- Local members of the evaluation team should be free to endorse the resulting final report or not, as they see fit.
- The evaluation team should minimise demands on interviewees' time.

NOTE: *Evaluations sometimes uncover evidence of wrongdoing. What should be reported, how and to whom are issues that should be carefully discussed with the evaluation manager.*

4.4 Ethics in consulting with local people

Consultation with local people in local contexts is essential in most evaluations. The design and implementation of evaluation consultation should take the following into account to ensure sensitive and successful interactions.

NOTE: *These precepts draw on well-established ethical codes and standards for research, notably the EC's* Horizon 2020 Ethics.

- **Informed consent**, ensuring that:
  - all potential respondents fully understand what their participation involves;
  - consent is given by the individual(s) concerned;
  - the scope of consent is absolutely clear, for example, community leaders giving consent for evaluators to enter the community, individuals giving consent to be interviewed;
  - participants know they can withdraw from the interview at any time without any implications for their access to development support or services in the future.
- **Transparency**, ensuring that participants:
  - fully understand the purpose of the evaluation, including how and where its findings will be made available;
  - can ask questions and seek clarification about the evaluation.
- **Confidentiality and anonymisation**, ensuring that:
  - participants understand that whatever they share will be non-attributable, unless their explicit permission is granted, and that the origin of information provided in the evaluation report cannot be deduced;
  - interview notes and other personal data are anonymised to ensure there is a negligible risk of someone being identified from the data, for example by using identifier codes rather than names when writing up interview notes.
- **Data protection**, respecting:
  - the legal obligation to comply with the EU's General Data Protection Regulation (GDPR, Regulation 2016/679), regardless of where the evaluation takes place within or outside the EU, by EU or non-EU citizens;
  - compliance with national or regional laws and regulations on data privacy and security, as appropriate.

SEE: Section 4.5 *for more detail on data privacy and security.*

- **Safeguarding**, in accordance with international/EU safeguarding standards for children, women, people with disabilities etc.

SEE: *The* EU Strategy on the Rights of the Child and the European Child Guarantee, *the* EU Gender Action Plan 2020–2025, *and the* Human rights and fundamental values *web pages.*

- **Non-discrimination** by evaluators when designing and carrying out an evaluation – for example, on the basis of race, gender, religious and non-religious convictions, sexual orientation, political affiliation, national origin, ethnicity, language, age or disability – while recognising that evaluation methods usually warrant appropriate disaggregation and representation of population

groups so different experiences and perspectives can be captured.

- **Reciprocity** in relationships and exchange of information with participants, ensuring that evaluation commissioners and evaluators:
  - fully recognise and value participants' contributions in terms of information, knowledge, perspectives and time and thus ensuring their time is not wasted;
  - commission evaluations with the intention of helping organisations effectively serve the needs of participants.

NOTE: *The Aotearoa New Zealand Evaluation Association is possibly the only authority including the principle of reciprocity in its evaluation standards (ANZEA, 2015).*

Evaluators must **respect the customs, beliefs and values** of the communities and countries in which they are operating (Aronsson and Hassnain, 2019). This has particular implications when designing and carrying out an evaluation, especially in complex environments. Box 4.1 provides an example of how this is relevant; Box 4.2 details an interesting EC initiative advocating for interculturalism.

NOTE: *While evaluation team members are expected to respect other cultures, they must also uphold EU values, especially with regard to minorities and women. The United Nations Universal Declaration of Human Rights is the operative guide in such matters.*

BOX 4.1 A practical example of respecting local culture, customs and beliefs

Customary protocol in many rural communities means the evaluation team should first meet and consult with the community leader, who is often male, to explain the purpose of the evaluation and the length of time the evaluation team will be in the community.

They should discuss any constraining factors (e.g. if the evaluation takes place during harvest time and farming members of the community have limited availability to talk to evaluators) and determine necessary adaptations (e.g. short interviews with farmers in the evening when they return from the fields).

The evaluation team should arrange to interview different groups within the community fully respecting culture and customs. For example, it may be appropriate for female evaluators to carry out all interviews with women in the community in a private location (e.g. the female quarters of a private house or a safe community space); this practice may also ensure evaluators hear views and experience beyond those of community gatekeepers (e.g. male leaders).

When access to specific vulnerable target groups that are essential to an evaluation is not possible, this will have repercussions on the validity of the evaluation and its findings. This shall be duly acknowledged by evaluators in their reports.

4.5 Ethics in collecting and managing data

Collecting and managing data for an evaluation represents a nexus of legal and ethical considerations, with data protection at the forefront. In the EU, the key provisions in this respect are set down in the General Data Protection Regulation (GDPR, Regulation 2016/679). The regulation is particularly focused on personal data and defines seven key principles on which data protection should be based:

- **Lawfulness, fairness and transparency.** Data collection should be compliant with legal provisions of the GDPR and any other legal orders relevant for the organisations involved in the data collection process.
- **Purpose limitation.** Personal data should be collected only with a specific purpose that is clearly stated for those who are concerned.
- **Data minimisation.** Only personal data necessary for the objectives of an evaluation should be collected.

BOX 4.2 Ethics in different cultural contexts: the Intercultural Approach

The Directorate-General for International Partnerships' (DG INTPA's) Intercultural Approach (InCA) initiative aims at 'increasing awareness and methodological agility to value the complexity of culture as [a] strategic modality for a successful evaluation process and international partnerships relations'. Specifically, InCA explores how different cultural layers – including family, gender, workplace, religion, education, ethnic, nationality and professional roles – affect us both individually and professionally, in our communities, work teams and organisations. By adopting wider and multiple frames of reference, we obtain enriching benefits brought by an intercultural perspective, values and ways of working. Treating every participant with respect allows each to feel seen and empowered, thus helping prevent and minimise any incidents or misunderstandings cultural differences could have on any successful, current or future, intercultural collaboration.

In evaluation processes specifically, the InCA ethical framework actively promotes cultural humility across working environments and conditions. This means all the different cultural layers of diversity (gender, nationality etc.) of all the involved groups/actors (EU delegations and external partners, stakeholders, beneficiaries, experts etc.) are considered through the lenses of different ethical principles: respect, empowerment, protection, responsibility, commitment to relationship.

Because cultural influences can significantly affect the evaluation process, the intercultural ethical framework facilitates the recognition of (and a commitment to) culturally responsive and collaborative evaluation approaches. Through the intercultural evaluation lens, participants actively co-create a respectful evaluation. When all involved can voice their needs, ideas, visions and values and feel heard, real data, clear information and freedom of expression emerge. This will also support the sustainability of any future evaluation of the partnership.

InCA activities aim to build an appreciation of how cultural diversity can influence ways of working, observing and evaluating. Specific suggestions are provided to adopt a deeper intercultural approach to value the different cultural layers for a more strategic and efficient result/evaluation. For more information, see the InterCultural Approach Programme (InCA) - Learning&Sharing Community on the Capacity4dev website.

- **Accuracy.** Data of individuals need to be kept accurate, and modified or deleted upon individuals' request.
- **Storage limitation.** Personal data should be deleted after they are no longer necessary.
- **Integrity and confidentiality.** The GDPR does not define these precisely, but data should be kept in line with the best available technologies and capacities of organisations.
- **Accountability.** Upon request, documents must be provided that demonstrate how the respective GDPR activities are executed.

In practice, when personal data are collected by evaluators, respondents should be made familiar with the purpose of the task and how their data will be processed and stored. Ideally, respondents should be asked for a consent statement, which includes a physical or virtual signature (which can be a full signature or a box ticking).

SEE: *The EC Principles of the GDPR web page for more information.*

Data disclosure risk is one of the main challenges with which evaluators are confronted. An efficient mechanism should be put in place to prevent situations in which personal and other sensitive data could 'leak out' and get into the hands of people whose benefits are not related to the evaluation process.

Many intruders seek opportunities to gain access – legally or illegally – to personal data, even if it is against the law. If an evaluation relies on personal data that can be of particular value for non-evaluation purposes, an advanced data protection system needs to be put in place. This system should include risk assessments and mitigation strategies, and investments in professional support for data security such as software or paid professionals.

4.6 Ethics to ensure equity-focused and gender-responsive evaluations

Equity-focused and gender-responsive evaluation incorporates principles of gender equality, rights and the empowerment of the most marginalised – especially women, children, the elderly and people with disabilities – to provide credible information about the extent to which an intervention has resulted in progress towards intended and/or unintended results regarding gender equality and women's empowerment.

Being gender aware and having gender competencies throughout the evaluation process is critical in all contexts, particularly complex settings, such as situations of conflict and fragility.

Ensuring that an evaluation's outcomes and process drive positive change towards gender equality empowers the involved stakeholders and can prevent further discrimination and exclusion. Evaluation design should be built on, and informed by, an understanding of gender dynamics and gender social norms within the evaluation context.

> **SEE:** Box 1.3, Box 3.2.1 *and discussion under* Subsection 3.1.2 *for more information on gender-responsive evaluation; also see EvalPartners' e-course on* Equity-Focused and Gender-Responsive Evaluations*; SaferWorld's training materials and toolkits on* Gender Analysis of Conflict*; and Garred et al. (2018).*

4.7 Ethics in situations of fragility, conflict and violence

Ethical issues are likely to arise more frequently – and can be particularly challenging – for evaluations conducted in fluid and volatile contexts. Evaluation in such contexts is not 'business as usual' but inherently political, and as such, has the potential to exacerbate tensions or put individuals or communities at risk if the human and institutional biases are not thought through in advance, or the exercise is not planned and carried out with extra care (Aronsson and Hassnain, 2021).

In such contexts, respect of basic evaluation principles becomes more difficult and requires a greater level of attention and sensitivity. Additional measures should be taken to alleviate these challenges.

> **EXAMPLE:** *Take measures to ensure that an evaluation conducted remotely in areas with limited connectivity does not exclude some target groups, which could lead to a bias in data collected that inadvertently favours the opinions of some groups over others.*

In conflict areas, it will be necessary to ensure a higher level of sensitivity towards the contextual dynamics, and to make sure that an **evaluation does not exacerbate tensions or put individuals and communities at greater risk** (see Box 4.3) This is particularly important in unstable and volatile environments, but is also relevant in other contexts where discrimination exists.

> **NOTE:** *The evaluation should seek to integrate sensitive and timely conflict analysis, and gender analysis of conflict, in its design, approach, reporting and dissemination. This analysis should be aimed at understanding the background, history and causes of the conflict; identifying all relevant groups involved and their different perspectives; and, especially, identifying the drivers of conflict the evaluation could affect – and thus have the potential to exacerbate or escalate conflicts or, conversely, reduce future conflict and its risks. For more about conflict sensitivity, see the Capacity4dev* Resilience, Conflict Sensitivity and Peace *web page; also see ICE (2020).*

The guiding principle is '**do no harm**', which means that people must not be exposed to further harm as a result of the evaluation. Harm includes violence, rights abuses and/or physical hazards. Three key considerations need to be taken into account.

- **Could the design of the evaluation process reinforce divisions and inflame conflict?** For example, a poorly designed and facilitated focus group that brings hostile members of a community together could lead to a heated discussion – which

BOX 4.3 Addressing protection issues in conflict situations

When evaluators engage and interact with a group suffering oppression and abuse in a context of active and violent conflict, it could exacerbate the protection risks faced by that group. Group members may be interrogated by their oppressors and forced to disclose what was said once the evaluators leave, and/or be subject to further intimidation and abuse for having talked to outsiders. Evaluators must therefore:

- have a **strong understanding of the local context** before starting work, and especially of power dynamics and patterns of abuse and oppression;
- explore whether it is possible to **interview members, especially those of oppressed groups, in a safe environment**, thus minimising any protection risks associated with the evaluation process and outcomes;
- **not engage with such groups if doing so might increase the risks they face**.

For example, in one context of protracted conflict, a group of displaced people of a certain ethnicity was living in a host village populated by another ethnic group. The displaced people were being forced to share a large proportion of their harvest and to provide agricultural labour on exploitative terms. A member of a national non-governmental organisation of the same ethnicity as the displaced group liaised with the evaluation team to facilitate discussions and build trust. A group interview was held away from any public space in the village. When some resident villagers joined the group, the facilitator, recognising the change in dynamic, stopped asking questions about protection issues which would have put the displaced group at greater risk.

not only could not be resolved in the context of the focus group but could trigger aggression and even violence within the community once the evaluators have left (or even while they are still present).

- **Could participating in the evaluation put population groups and/or individuals at risk of further harm?** Box 4.3 provides an example along with an appropriate mitigating action. A less political example of putting participants at risk is where evaluators ask community leaders to travel for a meeting during the rainy season, thus exposing them to physical dangers such as flooding or landslides.
- **Could the publication or sharing of evaluation findings put population groups and/or individuals at risk of further harm?** For example, when the views of individuals and groups most at risk are exposed insensitively, it could leave them vulnerable to reprisals by powerful actors (ALNAP, 2018). At the other end of the spectrum, when the views of a particular population group have been overlooked and not included, the evaluation's resulting analysis is biased – which could inform future allocation of aid resources and thus endanger that group's access to potentially life-saving assistance.

NOTE: *The GDPR defines specific rules for exposing the views of individuals, including their expressed prior approval.*

In each of these examples, a lack of awareness of the local context, negligence or careless design and implementation of the evaluation process by the evaluators could inadvertently cause harm. Three key measures can help avoid this:

- For all evaluations, assess whether any steps in the evaluation process could contribute to tensions and take mitigating action accordingly.
- For conflict settings, carry out a new (or update an existing) conflict analysis, to inform the planning and design of an evaluation.
- Revise the evaluation plans in light of the conflict analysis to ensure that they do not contribute to tensions and do harm.

NOTE: *These steps are drawn from Buchanan-Smith, Cosgrave and Warner (2016).*

Annex: Budget support

Intervention logic for budget support

It could be said that the evaluation of budget support has a strong heuristic dimension, compared to the normative dimension of most evaluations: it needs to find new narratives to explain successes and failures rather than just analyse whether and how successfully the predefined narrative has materialised.

The intervention logic underpinning budget support recognises that it is a contribution to the implementation of given policies and public spending actions of a partner government, according to a **results chain comprising five levels** rather than the standard four applied in the project modality. In the case of budget support, an additional induced outputs level is added to take account of the changes induced by inputs and outputs at the policy and institutional level, but not yet at the final beneficiary level (outcomes). By way of example, according to current definitions from the Development Assistance Committee of the Organisation for Economic Co-operation and Development, the legal and institutional accomplishment of a policy reform has to be considered as an output in the intervention logic. It cannot be considered as an outcome, because it does not per se represent a change in behaviour or a benefit to the people targeted by the programme. On the other hand, accomplishment of a policy reform is not a direct output of budget support, but rather an accomplishment of national stakeholders influenced by a number of factors, of which budget support is only one. That is why this crucial, additional level of the results chain (induced outputs) is introduced for budget support.

The five levels of a budget support results chain follow.

- **Level 1: Inputs.** These include funds, dialogue and technical assistance/capacity development support.
- **Level 2: Direct outputs.** These are the opportunities created by the deployment of the inputs. Typical direct outputs would include:
  - increased size and share of the **government budget** available for discretionary spending;

- improved **policy dialogue** framework (identification of modalities, instances and actors involved at the different levels), alignment of the performance indicators and related monitoring;
- better coordinated **capacity development** support, consistent with government priorities, sufficiently flexible, and conducive to the effective implementation of government strategies;
- improved **external assistance coordination, harmonisation and alignment** with the government's policies and implementation systems, including a reduction in the transaction costs of providing and receiving external assistance.

- **Level 3: Induced outputs.** These are the actual changes emanating (inter alia) from the use, by the partner government, of the opportunities created by budget support. Typical induced outputs would include:

  - improved **macroeconomic and budget management** (e.g. improved revenue and expenditure policies, inflation and debt management, monetary and foreign exchange policies etc.);
  - strengthened **public financial management and procurement systems** (e.g. improved fiscal discipline, transparency and oversight, enhanced allocative and operational efficiency, and better policy coordination);
  - strengthened **public sector institutions** in the targeted areas (e.g. improved planning capacities, monitoring systems, sector coordination);
  - improved **sector policies and implementation plans** (e.g. improved legal frameworks, better plans of action for reform implementation, implementation of key reform steps);
  - other improvements in **governance** (e.g. measures put in place to tackle corruption, improved local development policies);
  - increased quantity and quality of **goods** (such as public infrastructure) **and services** made available by the public sector.

- **Level 4: Outcomes.** These are the short- and medium-term changes resulting (inter alia) from the induced outputs. Typical outcomes would include:

  - increased and more equitable **use of the goods and services** made available by the public sector in the areas targeted by government policies and actions supported by budget support;
  - higher **levels of satisfaction** (perceived and actual benefits) of service users in the different areas targeted by budget support (education, health, social protection, justice);
  - improved **economic conditions** for different actors, such as employability, jobs created, business confidence and growth of private sector investment;
  - signs of improved **competitiveness** of the economy (e.g. increased foreign direct investment);
  - signs of improved **gender** equity and equality;
  - improved **citizen confidence** in the performance and transparency of the government, particularly regarding basic freedoms, public financial management and service delivery.

- **Level 5: Impact.** These are the higher-level results to which budget support will contribute depending on the targeted sectors. Typical impacts would include:

  - enhanced **socially and environmentally sustainable growth**, including employment, poverty reduction, protection of natural resources and mitigation of climate change;
  - enhanced **good governance**, peace, transparency, equitable justice and consolidated respect of fundamental rights;
  - enhanced **gender equity and social equality**.

There will of course be other impact areas to consider, depending on the specific partnership framework (i.e. accession, association or development cooperation) and the related priorities established. Intermediate impacts may also be considered, if necessary, to better reflect such priorities.

In the case of Directorate-General for International Partnerships (DG INTPA) evaluations, an indicative evaluation design is developed as part of the technical offer submitted by tenderers which is then detailed and finalised during the inception phase.

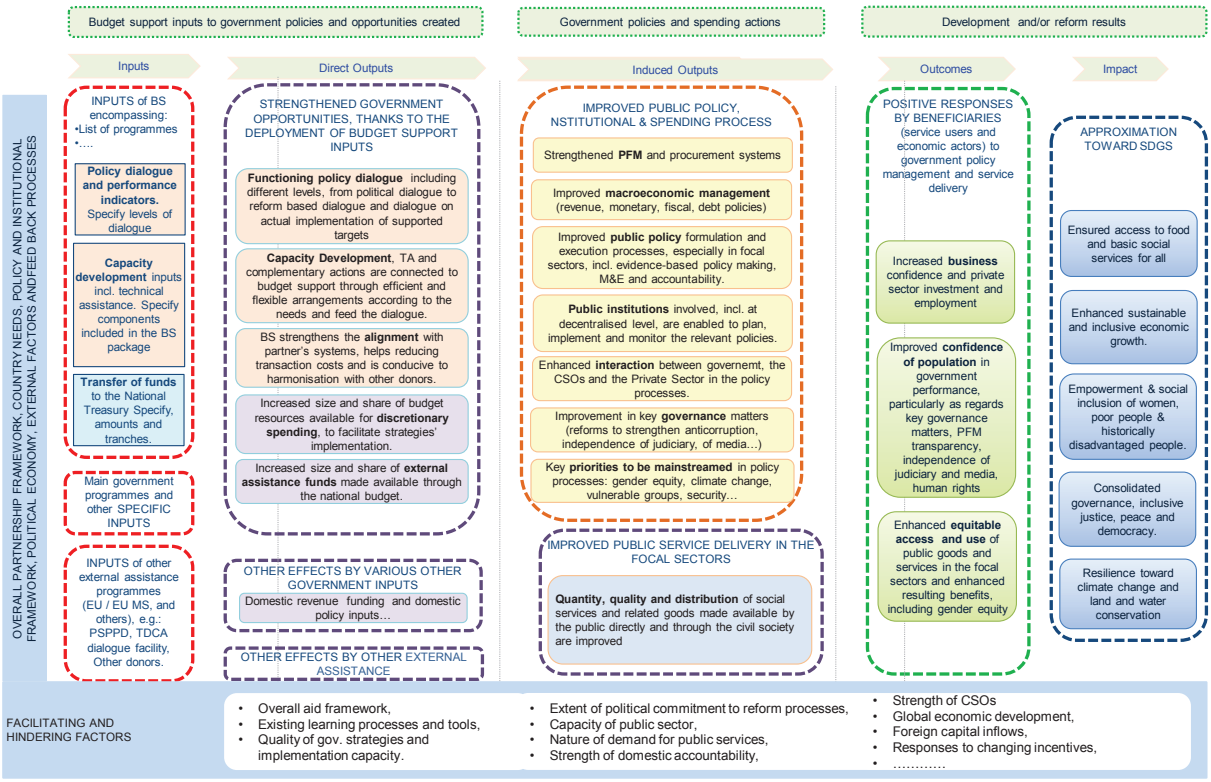
Methodology for evaluation of budget support

The intervention logic for budget support evaluations (see Figure A.1) explicitly recognises that the results of budget support are influenced by a variety of actors and factors – most notably government policies, measures and spending actions (which may or may not have been supported by budget support or by other modalities), civil society and private sector initiatives, as well as other exogenous factors (e.g. commodity prices on the world market, external capital inflows, political [in]stability).

In most cases, it will be possible to trace the specific contribution of budget support up to the level of induced outputs (Level 3), as the main processes and actors are clearly identifiable and the influence of budget support is relatively direct. Such direct tracing is not possible at Level 4 (outcomes) and Level 5 (impact), because budget support influence on these levels is indirect or marginal and many other, even unknown, actors and factors intervene in their determination.

Therefore, carrying out an analysis based on the assumption of direct causality between budget support inputs and direct outputs on the one hand, and outcomes and impact on the other, will not be fruitful, and in many cases will be impossible. Furthermore, such an analysis could lead to underestimating or overlooking non-budget-support-related policy or non-policy factors, which might have played a (major) role in the achievement of outcomes and impact. This acknowledgement of the limits of direct causality analysis in the case of budget support gave rise to the development of the **three-step approach**.

FIGURE A.1 Budget support intervention logic



SOURCE: Caputo (2022).

The three-step approach recognises that:

- **The contribution of budget support can be (more or less), credibly traced from Level 1 to Level 3**, though this does not mean that the influence on induced outputs (Level 3) of various actors and factors that are not part of budget support arrangements and packages should be disregarded. The influence of non–budget support factors is particularly important in more complex reform processes, where it may even be stronger than budget support effects. In the case of budget support to justice reform, for instance, it should not be difficult to identify a causal relationship between the achievement of certain reform targets (such as a new legislative framework, or the establishment of a formally independent council of the judiciary) and the contribution of budget support through intensive technical assistance and dialogue actions or increased budget availability for the sector. However, this does not mean that other factors were not at least as important, such as the determination of political parties in parliament, pressure from civil society and the free press, and/or the mobilisation of judges.
- **At outcome and impact levels (Levels 4 and 5), the specific contributions made by budget support are very difficult to trace through a linear causality assessment.** In the first place, government policies affecting results at these levels are manifold and are influenced by many non–budget support factors. Secondly, context-related factors – including the political economy framework, the international economic and political environment etc. – also play a strong determining role. On the other hand, at Level 4, it is generally possible to trace the role of government policies because the outcomes are often a response to the changes made by such policies, and therefore the establishment of causal relationships may be justified (though again, other factors must be considered). The same may apply to intermediate impacts, while longer-term impacts (Level 5) can only be estimated.

From the above, it can be concluded that it is generally possible to assess the contribution of budget support to induced outputs (Level 3), and the contribution of government policy to outcomes and impacts (Level 4 and partly Level 5). The issue is therefore how to combine these two contributions (budget support to induced outputs and induced outputs to outcomes and impact). The three-step approach provides a mechanism to do so (see Figure A.2).

STEP 1

Step 1 involves an assessment of budget support inputs, in relation to the context, the direct outputs generated by the inputs, and the induced outputs (Levels 1, 2 and 3 of the intervention logic). It also identifies the **plausible causal links** between these three levels and the role played by non–budget support factors in the chain of effects.

This assessment responds to the question of how, and to what extent, budget support has contributed to improving the quality and adequacy of government policies and service delivery systems for which the budget support partnership between the external partners and the government has been established.

> **NOTE:** *The evaluation does not assess the quality of the government actions as such, but their changes in view of achieving the objectives set out in the budget support contract.*

The assessment at induced output level should be extended to include consideration of additional external factors that may have influenced government policies other than budget support.

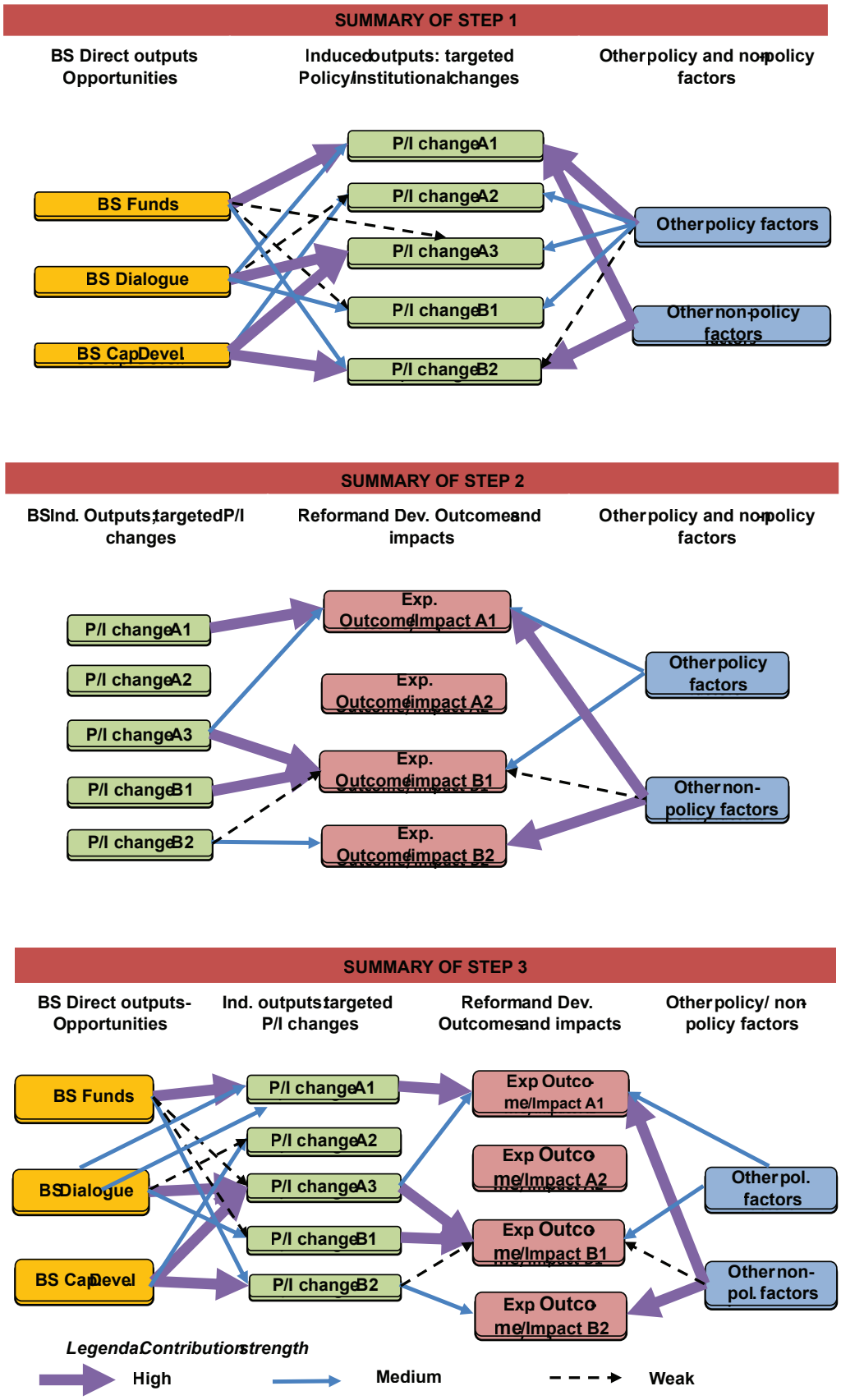
STEP 2

Step 2 entails assessment of the actual outcomes and impact(s) (Levels 4 and 5) as targeted by the government and supported by budget support. This step also includes an assessment of the **plausible causal links** between the government policies supported by budget support – among other factors – and those outcomes and impact(s).

There are two guiding questions for this step:

- **Were the achievements as regards the expected outcomes and (potential) impacts coherent with the original targets set by the government and supported by budget support?** A key issue here regards the time frame.

FIGURE A.2 Summary of the three-step budget support approach



SOURCE: Caputo (2022).

The evaluation will focus on the outcomes – if any – targeted by budget support during its life cycle, and will also assess the potential evolution of those outcomes and the impacts expected beyond the duration of budget support.

- **What were the main (positive or negative) causal factors of these achievements, including the specific role of the government's policies and strategies (as assessed at Level 3), the role of other actors and external factors?** Here as well, the time frame is of fundamental importance. In an intervention-level evaluation covering a rather short period of time (e.g. up to three years), it is not feasible to carry out a thorough assessment of outcomes and impacts including the related contribution analysis. Outcomes that can be recorded in short periods of time are either still weak, at an early stage, difficult to analyse in terms of causality (e.g. improvement of employability in a sector reform performance contract on vocational education and training), or are the clear consequence of easily identifiable events or actions (e.g. increased school enrolment among disadvantaged groups as a result of the construction of new classrooms in remote areas). In such cases, Part B of Step 2 does not take place. Step 2 is limited to Part A (as is explained more clearly later). On the other hand, in the case of strategic evaluations covering longer periods of time, thorough contribution analyses are necessary to respond to Part B of Step 2, to identify – by tracking retrospectively, towards the induced output level and other contextual factors – the factors that contributed most to the attainment of the observed outcomes and impact.

STEP 3

Step 3 is carried out by combining and comparing the results of Steps 1 and 2.

Step 3 is also based on a contribution analysis, although through an intermediate link; it establishes to what extent budget support contributed to the outcomes and impacts via its contribution to the supported government policies and actions. The crucial question is to what extent did the inputs of budget support and the opportunities they created (Levels 1 and 2) contribute to those government policies and strategies (Level 3), which in turn contributed to the actual outcomes and impacts observed at Levels 4 and 5. In most cases, it will be difficult to isolate the contribution of budget support, which means that the contribution has to be assessed through logical reasoning and by identifying logical linkages. Step 3 is also the occasion to verify and fine-tune the entire evaluation. Comparing the preliminary conclusions of Steps 1 and 2 allows their coherence and consistency to be verified. Indeed, insights obtained from Step 2 analyses might shed new light on the preliminary findings of Step 1, while drawing conclusions regarding budget support contributions (Step 3) might help to further specify the conclusions of Step 2. Finally, Step 3 might facilitate a better understanding of the role of external factors (economic, political and social contexts).

It is important to note that the three steps are in principle meant as analytical steps rather than chronological steps, with the obvious caveat that Steps 1 and 2 must cover comparable periods and must precede Step 3. Regardless of whether Steps 1 and 2 are carried out in parallel or in chronological order will depend on the organisation, data availability and requirements of a specific evaluation.

> **NOTE:** *For instance, Step 2 may use policy impact studies carried out before the evaluation has started, provided they cover the period of time under evaluation.*

SUMMARY

- **Step 1** is an evaluation of the effects of external support, which starts with the identification of the inputs, then tries to trace their contribution along the results chain towards the actual government policy outputs (i.e. the induced outputs). It entails an analysis of causal relations between well-identified inputs, and direct and induced outputs. It assesses to what extent budget support, based on a relevant design owned by the partner government, has contributed to strengthening government opportunities in terms of financial and policy capacities. It then assesses to what extent the new opportunities (together with other factors) have been used to strengthen policies, institutions, budget allocation processes, public financial management and service delivery.

- **Step 2** is a policy impact assessment which starts from the identification of the achievements in terms of the outcomes and impact(s) of the government policies targeted by budget support. Using various methods, and drawing on available resources, it aims to identify the factors that have contributed to the evolution of those outcomes and impact(s) over the evaluation period. This distinction calls for the use of specific and different methods such as existing policy impact assessment studies, surveys, regression analyses and other statistical tools.
- **Step 3** identifies the contribution of budget support to the expected/planned medium-term outcomes and impact by comparing, on the one hand, the contribution of budget support to government outputs and induced outputs, and on the other hand, the contribution of these to the expected/planned outcomes and impact.

The three-step approach has been fully applied in the evaluation of joint programmes, covering periods of 7–10 years and more, and involving large budget support providers, such as the European Union (EU), the World Bank, the African Development Bank, and some EU Member States. It was also applied to assess EU budget support, for instance in South Africa (2013) for a cluster of EU sector budget support interventions. More recently, it was applied to EU budget support in Paraguay (2016), Peru (2017), Cambodia (2018), El Salvador (2019) and Morocco (2021). In the context of IPA countries, the three-step approach is now being applied in Albania. While there are no previous applications of the comprehensive three-step approach, there have been a few intervention-level evaluations of budget support and the Evaluation of the Sector Approach under IPA II (2018), which was not a full evaluation of the specific contracts, but which has provided key information and produced important assessments on the SRPCs under way in the area.

In other cases (namely intervention-level evaluations), when the evaluation concerns one or more programmes over a short implementation period (approximately three years), or the scope of the evaluation is limited by the responsible bodies, the three-step approach is only partially applied. The intervention logic remains the same, because budget support works in the same way: preparation and design including the standard inputs and conditions → creation of opportunities for the partner government → improvement of policy and institutional framework → generation of outcomes and impact. The evaluator, however, is only expected to complete Step 1, while Step 2 will be limited to the assessment of actual and potential outcomes and impacts expected by the end of the budget support contract and afterwards and to a comparison with Step 1. The evaluator will therefore not have to assess the contribution of the policy and institutional changes supported by the budget support (i.e. the induced outputs) to these outcomes and impacts.

In this case, Step 2 will identify the actual level of outcomes and impacts, their evolution during the evaluation period and their potential for future development. In addition, it will assess the consistency between the dynamics of outcomes and impacts and the policy and institutional induced outputs assessed in Step 1.

Glossary

NOTE: *The following abbreviations and acronyms are used herein: EC = European Commission; EU = European Union; DG INTPA = Directorate-General for International Partnerships; OECD = Organisation for Economic Co-operation and Development.*

Accountability. Obligation to demonstrate that work has been conducted in compliance with agreed-upon rules and standards or to report fairly and accurately on performance results vis-à-vis mandated roles and/or plans. This may require a careful, even legally defensible, demonstration that the work is consistent with the contract terms.

Activity. Actions taken or work performed through which inputs, such as funds, technical assistance and other types of resources, are mobilised to produce specific outputs (OECD DAC, 2023).

Adaptive management. A structured management strategy that involves an ongoing process of working collaboratively and flexibly to learn, make decisions, test assumptions and adjust actions on the basis of new information, lessons and changes in context (OECD DAC, 2023).

Additionality. When (financial or non-financial) inputs, activities or results of a strategy, policy, instrument, modality, intervention or set of interventions are considered additional compared to what would have happened otherwise (OECD DAC, 2023). See EU added value.

Assumption. External necessary and positive conditions – not under intervention management or EU control – that must hold in order for the results chain to be valid. Assumptions should be formulated based on the context analysis and the risk assessment (*INTPA Companion*).

Attribution. The ascription of a causal link between observed (or expected to be observed) changes and a specific strategy, policy, instrument, modality, intervention or set of interventions. This does not require that changes be produced solely by the evaluand, but represents the extent to which observed effects can be attributed to it or to one or more partners, taking account of other interventions, confounding factors (other influences) or external shocks (OECD DAC, 2023).

Audit. Intended to detect and prevent irregularities in the implementation of inputs, processes and outputs based on criteria that are known and clarified in advance (e.g. budgets, regulations, management and technical standards). A financial audit focuses on compliance with applicable statutes and regulations; a performance audit focuses on relevance, economy, efficiency and effectiveness; and an internal audit assesses internal controls.

Baseline. Initial reference point or value for a given indicator that compares with actual values, allowing evaluand results to be assessed.

Baseline study. An analysis describing the situation prior to a development strategy, policy, instrument, modality, intervention or set of interventions, against which progress can be assessed or comparisons made.

Benchmark. A reference point or standard against which changes, performance or achievements can be assessed (OECD DAC, 2023).

Beneficiaries. The individuals, groups or organisations – whether targeted or not – that benefit, directly or indirectly, from a strategy, policy, instrument, modality, intervention or set of interventions (OECD DAC, 2023). Beneficiaries can be direct (those who benefit at first hand and in the short term) or indirect/final (those who benefit from an evalaund's outcome or impact in the long term) at the society or sector level. Targeted beneficiaries are those whose action or change in behaviour is sought through a particular strategy, policy, instrument, modality, intervention or set of interventions and thus are directly affected by it.

Bias. Conscious or unconscious prejudice in favour of or against a particular person, group or thing. Evaluators can be biased in the way they analyse data; users of evaluations can be biased in the way they interpret findings.

Big data. The use of data in large amounts, which is often characterised by a high degree of complexity. Big data may directly contribute numerically to an evaluation as well as provide a picture of a situation in which changes occur. For example, big data derived from social media can help measure public perception (citizen interest or trust) of an evaluand. Big data may raise ethical concerns, which need to be taken into account.

Case study. A detailed account giving information about the development of a person, group or thing, especially so as to illustrate a general principle.

Causal evaluation question. Evaluation question that asks about why results happened, especially the connection between a strategy, policy, instrument, modality, intervention or set of interventions and outcomes and impacts (intended or unintended).

Coherence. The compatibility of an evaluand with other interventions in a country, sector or institution (OECD DAC, 2023). One of the seven criteria used in evaluation by the EC.

Comparison group. A non-randomly selected group that does not receive the services, products or activities of the evaluand (USAID, 2009). Also see control group; treatment group.

Conclusion. Draws on data collection and analyses undertaken through a transparent chain of arguments to summarise factors leading to the success or failure of the evaluand (OECD DAC, 2023).

Conflict-sensitive evaluation. Incorporates a detailed understanding of the context in terms of historical, actual or potential conflict into traditional evaluation activities and its dissemination. Conflict-sensitive evaluations are used to understand the overall impact a given intervention has had on the context and that the context has had on the intervention (International Alert, 2004).

Context. The setting in which an intervention or evaluation takes place and which is likely to influence performance and results. These include capacities and social, economic, political, environmental, conflict, inclusiveness, cultural and institutional conditions (OECD DAC, 2023).

Control group. The sample or group that does not receive the intervention and against which other groups or samples (that do receive the intervention) are compared in order to assess performance and results (OECD DAC, 2023). Also see comparison group; treatment group.

Counterfactual. Situation or condition that hypothetically may prevail for individuals, organisations or groups where there is no intervention (the status quo) (OECD DAC, 2023). It may relate to the absence of the strategy, policy, instrument, modality, intervention or group of interventions under evaluation.

Data. Characteristic of information, which can be expressed in either qualitative or quantitative ways. Data are representations of variables. Further typologies concerning data are their origin (primary or secondary), mode of collection (in situ, ex situ, desk or field) and level of aggregation (raw or processed). Also see big data; knowledge.

Data collection method. Method used to identify information sources and collect information. Examples include informal and formal surveys, direct and participatory observations, community interviews, focus groups, expert opinions, case studies, and literature search (OECD DAC, 2023).

Data source. The location where information originates. Data sources should be relevant, trustworthy, attainable and regularly available. See primary data and secondary data.

Descriptive evaluation question. Evaluation question that asks what has happened and requires evaluators to define, observe and measure change, often from the point of view of various stakeholders. These questions pertain to positive and negative changes, be they expected or unexpected and directly or indirectly linked to the evaluand.

Effectiveness. The extent to which an evaluand achieved, or is expected to achieve, its objectives and results, including any differential results across groups (OECD DAC, 2023). One of the seven criteria used in evaluation by the EC.

Efficiency. The extent to which an evaluand delivers, or is likely to deliver, results in an economic and timely way (OECD DAC, 2023). One of the seven criteria used in evaluation by the EC.

Endline. The conditions existing after an intervention, or end of the period, against which changes from the baseline can be measured, monitored and evaluated. (OECD DAC, 2023) The value of an indicator at the end of an intervention.

Equity. The quality of impartiality and fairness; associated with strategies to reach an equal society.

Ethics. In evaluation, provides a framework and guidance to help evaluators practice ethically throughout the development, design, implementation, adaptation, presentation, interpretation and use of an evaluation (van den Berg, Hawkins and Stame, 2022a).

EU added value. The additional benefits created by the EU's (versus Member States) having carried out an action in a partner country. It directly stems from the principle of subsidiarity as defined in Article 5(3) of the Treaty on European Union. One of the seven criteria used in evaluation by the EC.

Evaluability. Extent to which an evaluand can be evaluated in a reliable and credible fashion. Some approaches to evaluability assessment involve early review to ascertain whether objectives are adequately defined and results are verifiable. In other instances, particularly with complex interventions, high uncertainty or in unstable contexts, evaluability assessment might instead identify a need for an evaluation approach that supports adaptive management (OECD DAC, 2023).

Evaluability assessment. Study conducted to determine (i) whether an evaluand is at a stage at which progress towards objectives is likely to be observable, (ii) whether and how an evaluation would be useful to managers and/or policymakers and (iii) the feasibility of conducting an evaluation (USAID, 2009).

Evaluand. The subject of an evaluation. This handbook uses the term generically to refer to any strategy, policy, instrument, modality, intervention or group of interventions assessed as part of EU development cooperation.

Evaluation. The systematic and objective assessment of a planned, ongoing or completed strategy, policy, instrument, modality, intervention or group of interventions, including its design, implementation and results. The aim is to determine relevance, coherence, effectiveness, efficiency, impact and sustainability. In the context of the EC, EU added value must also be assessed. Not all evaluations deal with all criteria or to the same degree. Evaluation also refers to the process of determining the worth or significance of an intervention.

An evaluation should provide information that is credible and useful, enabling the incorporation of lessons learned into decision-making processes (OECD DAC, 2023).

Evaluation approach. Comprises both the evaluation design and the evaluation methodology.

Evaluation design. Details the data sources, data collection processes and analysis methods used to answer the evaluation questions and associated judgement criteria.

Evaluation matrix. Mandatory format used in DG INTPA and FPI evaluation reporting to summarise the methodological design of, and documenting the evidence analysed for, each evaluation question.

Evaluation methodology. How the evaluation is conducted. Designing the methodology includes defining the judgement criteria and indicators for each evaluation question and selecting the data collection tools and sources to be used. The methodology should be gender sensitive, contemplate the use of sex- and age-disaggregated data, and assess if and how the evaluand has contributed to progress on gender equality.

Evaluation questions. High-level questions that the evaluation aims to answer; does not refer to questions in a data collection tool such as a questionnaire or interview guide. Typically, evaluation questions are either causal, descriptive or normative.

Evaluation tool. See data collection method.

Evidence. Facts or information that support the validity and truth of a conclusion, assumption or assertion (OECD DAC, 2023).

Evidence-based. Reliable and credible evidence determines the design, adaptation and implementation of a policy or practice (OECD DAC, 2023).

Ex ante evaluation. Performed before adopting or implementing an intervention. It supports intervention design and tests likely effects or scenarios to fine-tune the design or prepare for future evaluations.

Ex post evaluation. Occurs one to two years after operational closure of an intervention. This type of evaluation is mainly concerned with assessing the impacts generated by the intervention and verifying the sustainability of the accrued benefits.

Ex situ data collection. When data are collected using available documentation from books or websites, rather than undertaking fieldwork.

External validity. The degree to which the findings, conclusions and recommendations produced by an evaluation are applicable to other settings and contexts.

Final evaluation. Generally takes place shortly before or after operational closure of an intervention to contribute to accountability and learning and to identify any lessons that would help improve the quality of future interventions as well as the strategic decisions of the delegation/unit. Ideally, final evaluation occurs shortly before closure of an intervention to ensure the implementing team is still available and up to speed.

Finding. Uses evidence from one or more evaluations to allow for a factual statement (OECD DAC, 2023).

Focus group discussion. Brings together between 5 and 10 people to discuss topics related to the evaluation. A focus group is a relatively cost-efficient means of data collection; however, in order to succeed, it requires a skilled facilitator.

Formative evaluation. Evaluation intended to improve performance or to inform planning of a subsequent phase, often conducted during the implementation phase of the intervention. Formative evaluations may also be conducted for other reasons such as compliance, legal requirements or as part of a larger evaluation initiative (OECD DAC, 2023). More generally, formative evaluation is an ongoing process that allows feedback to be implemented during the intervention cycle.

Fragility. The combination of exposure to risk and insufficient coping capacities of a state, system and/or communities to manage, absorb or mitigate those risks. It occurs in a spectrum of intensity across six dimensions: economic, environmental, political, security, societal and human (OECD, 2023). The OECD's fragility framework, through its depiction of the balance of risks and coping capacities across these dimensions, helps inform an understanding of the drivers and consequences of fragility, including responses in fragile contexts.

Gender. The socially constructed roles associated with being male and female and the relations between women and men and girls and boys. Unlike sex, which is biologically determined, gender roles are learned and change over time and across cultures (UNEP, 2016).

Gender equality. When women and men are treated equally by ensuring they have the same rights, opportunities and responsibilities; equal access to public goods and services; and equal outcomes.

Gender mainstreaming. 'The process of assessing the implications for women and men of any planned action, including legislation, policies or programmes, in all areas and at all levels. It is a strategy for making women's as well as men's concerns and experiences an integral dimension of the design, implementation, monitoring and evaluation of policies and programmes in all political, economic and societal spheres, so that women and men benefit equally, and inequality is not perpetuated. The ultimate goal is gender equality' (UN ECOSOC, 1997).

Gender-responsive evaluation. Assesses changes to gendered power relationships and results from an evaluand, determining whether and how changes have occurred, and the effects of those changes. The approach to the evaluation ensures that the voices of people of different genders are incorporated throughout the process and in the methods used (OECD DAC, 2023).

Grant. A financial donation awarded by a contracting authority to a grant beneficiary. It is funded by the EU general budget or the European Development Fund.

Hypothesis. A set of (testable) ideas, beliefs and explanations about the relationship between an intervention and its effects in a given context (OECD DAC, 2023). Hypotheses are generally formulated during the desk activities of the interim phase of an evaluation together with the preliminary answers to the evaluation questions; they are then tested (validated or revised) during the field phase.

Impact. The extent to which an evaluand has generated or is expected to generate significant positive or negative, intended or unintended, higher-level effects (OECD DAC, 2023). In DG INTPA, impact is the long-term change to which the evaluand contributes at the country, regional and sector levels in terms of benefit to the population (*INTPA Companion*). One of the seven criteria used in evaluation by the EC. Impacts are captured by indicators; examples of impact indicators are percentage of the population living below the line of poverty and the mortality rate of children under age five.

Impact evaluation. Assesses the degree to which an evaluand meets its higher-level goals and identifies causal effects. Impact evaluations may use experimental, quasi-experimental and non-experimental approaches (OECD DAC, 2023). Also can refer to evaluations that use explicit counterfactual analysis to determine the effects (including outputs and outcomes) caused by an intervention.

Independence. As used in the context of evaluation, implies freedom from political influence and organisational pressure, full access to information and full autonomy in carrying out investigations and drawing conclusions.

Indicator. Quantitative or qualitative factor or variable of interest, related to the intervention and its results, or to the context in which an intervention takes place (OECD DAC, 2023). A variable specifying how performance can be measured and assessed. Together with the results chain, indicators form the basis of an intervention's monitoring and evaluation system. For examples, see impact, outcome and output.

Information. Fact based upon evidence. In evaluation, this evidence is supplied with the collected data. Also see knowledge.

Input. Financial, human or material (in-kind), and institutional (including technological and information) resources used to conduct or carry out a strategy, policy, instrument, modality or intervention (OECD DAC, 2023).

In situ data collection. When data are collected via fieldwork or at the location of the evaluand.

Intervention. A coherent set of activities and results structured in a logical framework aimed at delivering development change or progress. An intervention is the most effective (and hence optimal) unit for

operational follow-up by the EC of its external development operations; it is thus used as the base unit for managing operational implementations, assessing performance, monitoring, evaluation, internal and external communication, reporting and aggregation. 'Intervention' is used in this handbook to refer generically to a project or programme.

Intervention-level evaluation. Analyses the results of a specific intervention or group of logically interlinked interventions within the frame of a wider scope of collaboration with a country or region. An integral part of intervention cycle management as it helps enhance the programming, design, implementation, performance and achievement of results of EU interventions.

Intervention logic. The way an intervention is expected to achieve its desired results, including underlying assumptions about the causality and interaction between the intervention, its inputs, activities, outputs, outcomes and impacts, in the context of the intervention (OECD DAC, 2023). Also see logical framework approach.

Joint evaluation. One carried out, in whole or part, by more than one organisation, and/or evaluation of interventions being implemented by more than one organisation. The EU encourages joint evaluations undertaken with partners and other donors as their involvement (i) aligns with aid effectiveness priorities and (ii) delivers on the EU commitment to increase joint programming and joint interventions, notably when funds are pooled (as in budget support, blending etc.).

Judgement criteria. Specify aspects of the evaluand that will allow its merits or success to be assessed. Two or more judgement criteria are derived for each evaluation question. . Each judgement criterion should be formulated as a positive statement and be used to answer an evaluation question positively or negatively and be accompanied by one or more indicators.

Knowledge. Gained through growing familiarity with data and information generated via the evaluation process, resulting in an understanding of why and how things happened, why in this and not another way, and what the most optimal solutions are to given problems.

Limitations. Any constraints in the process, methodology or data that affect monitoring or evaluation, including potential implications for validity and reliability (OECD DAC, 2023).

Logical framework approach. Systematic process to build the intervention logic, making it explicit and using analytical and planning tools that improve its design and allow for its relevant, feasible and effective outcome-focused management. Creates an intervention logic supported by three interdependent pillars: results chain (inputs, activities, outputs, outcomes, impacts), evidence-based assumptions (operational, behavioural, political) and monitoring system (indicators, baselines, targets, sources of verification). In DG INTPA and FPI, the logical framework approach is used at each stage of the intervention cycle to help ensure results-oriented delivery. The logical framework matrix is a product of the logical framework approach.

Logical framework matrix (logframe). A management tool used to improve the design of interventions, most often at the project level. It involves identifying strategic elements (inputs, activities, outputs, outcomes, impacts) and their causal relationships, as well as indicators, and the assumptions or risks that may influence success and failure. Facilitates planning, execution, monitoring and evaluation of an intervention (OECD DAC, 2023). Its use is mandatory in EC intervention design, monitoring and evaluation.

Meta-evaluation. A systematic and objective assessment that aggregates findings and recommendations from a series of evaluations.

Midterm evaluation. Performed about halfway during implementation of an intervention. It focuses on progress to date and explains why progress is – or is not – occurring as planned. It also provides recommendations on how to improve the intervention during its remaining duration in order to achieve expected objectives, taking into account both problems and opportunities and lessons for future interventions in the same sector or region.

Mixed methods. Use of both quantitative and qualitative methods of data analysis in an evaluation.

Monitoring. Continuing function that uses systematic collection of data on specified indicators to provide

management and main stakeholders of an ongoing development intervention with indications of the extent of progress and achievement of objectives and progress in the use of allocated funds. Ongoing analysis of intervention progress in achieving the expected results. Verifies the sound management of interventions and provides information on the use of inputs, implementation of activities, and progress in delivering outputs and achieving outcomes. Unlike an audit or evaluation, monitoring produces systematic information with a short periodicity. It should quickly detect discrepancies and direct implementation towards corrective actions. Also see results-oriented monitoring.

Normative evaluation question. Evaluation question that defines or sets possibilities or opinions, such as 'What should be the temperature in the room?' or 'Is the intervention worth the cost?' Also called valuing question.

Observation. Collecting data while studying human behaviours. Observation is a good means of cross-checking/validating information from other sources.

OPSYS (Operational System). Web-based information technology (IT) ecosystem for EC staff and implementing partners that will incorporate and replace all pre-existing IT systems for the management of the entire EU external cooperation portfolio.

Outcome. A short- to medium-term change in the behaviour of the target groups and/or effects on the political, social, economic and/or environmental areas targeted by EU action; the action will contribute to change at this level (it is under its influence but not direct control) (*INTPA Companion*). In budget support, refers to positive response by beneficiaries (service users and economic actors) to government policy management and service (EC, 2018a).

Outcome mapping. Methodology for planning, monitoring and evaluating development initiatives. An alternative to the logical framework approach.

Output. Direct deliverables or benefits of activities under the direct control of the action (*INTPA Companion*). The product, capital good or service that results from an intervention. Outputs may also include changes resulting from the intervention that contribute to the achievement of outcomes. Outputs include changes in knowledge, skills or abilities produced by the activities (OECD DAC, 2023). Outputs are captured by indicators; examples of output indicators are number of secondary schools equipped with technical vocational facilities as per agreed standards, number of teachers trained passing performance assessment (test) and number of kilometres of road rehabilitated.

Participatory evaluation. Evaluation method in which representatives of agencies and stakeholders (including beneficiaries) work together in designing, carrying out and interpreting an evaluation.

Primary data. Original data newly collected for the purposes of evaluation. Primary sources provide direct, first-hand evidence about the subject of the research. Primary evidence is gathered by evaluators through direct, qualitative and/or quantitative analysis using evaluation tools. Also see secondary data.

Programme. Temporary organisational set-up to manage a set of projects with a common goal and to obtain results and control not obtainable from managing them individually. In DG INTPA and FPI, often referred to as an intervention.

Project. Temporary set of coordinated activities to create a unique output within certain constraints such as time, cost and quality. In DG INTPA and FPI, often referred to as an intervention. Also previously used to refer to any implementing modality that is not budget support.

Proxy. Use of observation from which a fact in issue can be inferred when data cannot be obtained directly.

Quality assurance. Any activity or process used to assess and improve the merit or worth of an evaluand (including the quality of the evaluation deliverables) or its compliance with given standards and requirements (OECD DAC, 2023).

Qualitative data. Non-numerical data that characterise an object or event. Also known as categorical data since these data can be organised as a set of properties. In evaluation, qualitative data provide an opportunity to gain a rich picture that can help explain why and how something happened because of the evaluand.

Quantitative data. Numerical data collected to answer questions such as 'How many?' or 'How much?' in attempting to measure various qualities of an evaluand.

Quasi-experimental design. A methodology in which research subjects are assigned to treatment and comparison groups typically through some sort of matching strategy that attempts to minimise the differences between the two groups in order to approximate random assignment.

Randomised control trial. Research study that uses an experimental design.

Recommendations. Proposals aimed at enhancing the relevance, coherence, effectiveness, efficiency, impact, sustainability or EU added value of the evaluand; at redesigning the objectives; or reallocating resources. Recommendations should be based on findings and conclusions.

Reference group. Presided over by the evaluation manager and composed of colleagues and stakeholders, as well as representatives from partner countries and/or other organisations. Provides expertise, access, monitoring support, supervision and facilitation for the evaluation manager and evaluation team throughout the life of the evaluation.

Relevance. The extent to which the objectives and design of an evaluand respond to beneficiaries, global, country and partner/institution needs, policies and priorities, and continue to do so if circumstances change (OECD DAC, 2023). One of the seven criteria used in evaluation by the EC.

Result. The outputs, outcomes or impacts (intended or unintended, positive or negative) of an intervention (OECD DAC, 2023). Each element contributes to the next as set out in a results chain.

Results-oriented monitoring (ROM). External monitoring system reinforcing results-based management in EU external action operations as part of the EC's aid effectiveness and accountability. The ROM system supports and complements the monitoring and reporting activities carried out by the EU operational units managing the external action interventions' portfolio.

Results chain. The causal sequence of an intervention that stipulates the different stages leading to achievement of the desired objectives. In general, the results chain starts with inputs, which then link to activities and outputs, and culminate in outcomes and impacts (OECD DAC, 2023).

Review. Assessment of the performance of a strategy, policy, instrument, modality, intervention or set of interventions, periodically or on an ad hoc basis. Not the same as evaluation, which is generally more systematic and comprehensive. Reviews tend to emphasise operational aspects (OECD DAC, 2023).

Risk. Any uncertain event or set of events that, if realised, will negatively affect the achievement of the objectives and expected results set out in the intervention logframe. Lost opportunities are also considered to be risks (EC, 2018b).

Risk management. A continuous, proactive and systematic process of identifying, assessing and supervising risks in line with accepted risk levels, carried out at every level of the EC to provide reasonable assurance regarding the achievement of objectives. The five steps of a risk management process are (i) identification of objectives and outputs, (ii) risk identification and assessment, (iii) selection of risk response, (iv) implementation of risk response and (v) monitoring and reporting (EC, 2018b).

Rubric. A framework that sets out criteria and standards for different levels of performance and describes what performance would look like at each level.

Sample. Subset of a given population that is chosen so as to allow for extrapolation of findings to the full population (OECD DAC, 2023).

Secondary data. Material or evidence collected by someone other than the primary user. Common sources of secondary data include censuses, information collected by government departments, organisational records and data originally collected for other research purposes. Also see primary data.

Stakeholders. Agencies, organisations, groups or individuals who have a direct or indirect interest in the intervention or its monitoring and evaluation

(OECD DAC, 2023), including primary intended users and others. Different stakeholders can be engaged for different purposes and at different phases of evaluation planning and implementation.

Strategic evaluation. Assesses the results of EU strategies from conception to implementation at any or all of several levels – country, region, theme or sector policy or financing instrument – over an extended period of time (often 7–10 years).

Summative evaluation. Examines intervention outcomes to determine overall intervention effectiveness.

Sustainability. Extent to which the net benefits of an evaluand continue or are likely to continue (OECD DAC, 2023). One of the seven criteria used in evaluation by the EC.

Systematic review. An approach to synthesising evidence from multiple studies. Systematic reviews use methodical approaches and criteria to identify relevant studies for inclusion, assess their quality, extract data and synthesise evidence (Better Evaluation website, Systematic review web page).

Target. An objective, usually quantitative, defined as a value of an established indicator. The target is generally set at the beginning of an intervention and is expected to be achieved by a specific point in time with available resources (OECD DAC, 2023). Example: a 20 per cent increase in the rate of school completion for girls by the end of the intervention.

Target group. The specific individuals, communities or organisations that the intervention is intended to reach. Can also be defined as the recipients of the goods and services produced by the intervention, or whose skills or capacities have changed because of the intervention. The target group may or may not be the individuals or organisations that, ultimately, are intended to benefit from the intervention (OECD DAC, 2023).

Terms of reference (ToR). Document capturing the evaluation mandate for the evaluation team that will carry it out, including the context of the evaluation, the work to be performed by the evaluators and its structuring, any specific methodological requirements, any necessary expertise required of the evaluation team, key deadlines and deliverables and further elements.

Theory of change. The way an intervention is expected to achieve or achieves change. It represents how people understand change to occur in a given context, including explicit (or implicit) assumptions about the causal links between inputs, activities and results. Often also includes evidence and risks for these elements of the results chain (OECD DAC, 2023). In the EC, generally referred to as the intervention logic (EC, 2023).

Transparency. An environment in which the objectives of policy, its legal, institutional and economic framework, policy decisions and their rationale, data and information related to monetary and financial policies, and the terms of agencies' accountability are provided to the public in a comprehensible, accessible and timely manner (IMF, 1999).

Treatment group. Group that receives the services, products or activities of the evaluand. Also see control group.

Triangulation. Simultaneous use of multiple theories, sources or types of information, or types of analysis to verify and substantiate an assessment. Triangulation increases confidence in evaluation findings by basing them on various points of view.

Utility. A core premise of evaluation design and conduct to ensure that the evaluation and its recommendations are useful in the sense of recognising 'how real people in the real world apply evaluation findings and experience and learn from the evaluation process' (Patton 2013, p. 1).

Values-based evaluation. An approach to evaluation based primarily on the values of the evaluand. Values-based evaluation uses participation and cultural sensitivity, among other tools, to understand the deeply rooted values in a community or institution (Aronsson and Hassnain, 2019).

Variable. An attribute or characteristic of interest to be measured, recorded and analysed in an individual, group or system.

References

ANZEA (Aotearoa New Zealand Evaluation Association) (2015), 'Evaluation Standards for Aotearoa New Zealand', Superu, Wellington.

AES (Australasian Evaluation Society) (2013), 'Guidelines for the Ethical Conduct of Evaluations'.

ALNAP (2018), 'Evaluation of Protection in Humanitarian Action', ALNAP/ODI, London.

Aronsson, Inga-Lill, and Hur Hassnain (2019), 'Value-based Evaluations for Transformative Change', in Rob D. van den Berg, Cristina Magro and Silvia Salinas Mulder, eds., *Evaluation for Transformational Change*, chapter 6, International Development Evaluation Association (IDEAS), Exeter, UK.

Aronsson, Inga-Lill, and Hur Hassnain (2021), 'Evaluation and Ethics in Contexts of Fragility, Conflict and Violence', in Rob D. van den Berg, Penny Hawkins and Nicoletta Stame, eds., *Ethics for Evaluation: Beyond 'Doing No Harm' to 'Tackling Bad' and 'Doing Good'*, chapter 10, Routledge, New York.

Austrian Development Agency (2022), 'Evaluability Assessments in Austrian Development Cooperation: Guidance Document', Austrian Development Agency, Vienna.

BenYishay, A., D. Runfola, R. Trichler, C. Dolan, S. Goodman, B. Parks and A. Anand (2017), 'A Primer on Geospatial Impact Evaluation Methods, Tools, and Applications', Working Paper 44, AidData, William and Mary College, Williamsburg.

Bevans, Rebecca (2020), 'Choosing the Right Statistical Test: Types & Examples', Scribbr, blog post, updated 28 November.

Buchanan-Smith, M., J. Cosgrave and A. Warner (2016), 'Evaluation of Humanitarian Action Guide', ALNAP, London.

Caputo, Enzo (2022), 'Review of the Methodological Approach for Budget Support Evaluations', European Commission, Belgium.

CIDA (Canadian International Development Agency) (2001) 'How to Perform Evaluations: Participatory Evaluation', Quebec.

CIRAD (2015), 'Principles and Tools', on ImpresS (Impact of Research in the South) website.

Collier, D. (2011), 'Understanding Process Tracing', *Political Science and Politics*, 44 (04): 823–830.

Conflict Sensitivity Consortium (2012), 'How to Guide to Conflict Sensitivity', Conflict Sensitivity Consortium, London.

Connell, J. P., and A. C. Kubisch (1998), 'Applying a Theory of Change Approach to the Evaluation of Comprehensive Community Initiatives: Progress, Prospects, and Problems', *New Approaches to Evaluating Community Initiatives*, 2 (15-44): 1–16.

Corral, P., A. Irwin, N. Krishnan, D. G. Mahler and T. Vishwanath (2020), *Fragility and Conflict: On the Front Lines of the Fight against Poverty*, World Bank, Washington, DC.

Davidson, E. J., and T. K. Chianca (2020), 'Retrospective Impact Evaluation: Save the Children's Sponsorship Programming in Woliso Impact Area, Ethiopia (2002-2010)', Real Evaluation, Seattle and Rio de Janeiro.

Davies, R. (2018), 'Representing Theories of Change: Technical Challenges with Evaluation Consequences', CEDIL Inception Paper 15, London.

Davies, R. (2013), 'Planning Evaluability Assessments: A Synthesis of the Literature with Recommendations', Working Paper 40, Department for International Development, UK.

Davies, R., and J. Dart (2005), 'The "Most Significant Change" (MSC) Technique: A Guide to Its Use'.

DG INTPA (Directorate-General for International Partnerships) (2020), 'Creative Communications for Evaluation Dissemination', European Commission, Brussels.

DG INTPA (Directorate-General for International Partnerships) (2021), 'EvalCrisis Lessons Learnt Paper', European Commission, Brussels.

DG NEAR (Directorate-General for Neighbourhood and Enlargement Negotiations) (2016), 'Guidelines on Linking Planning/Programming, Monitoring and Evaluation', European Commission.

Douthwaite, B., T. Kuby, E. van de Fliert and S. Schulz (2003), 'Impact Pathway Evaluation: An Approach for Achieving and Attributing Impact in Complex Systems', *Agricultural Systems* 78 (2): 243–265.

Earl, Sarah, Fred Carden and Tom Smutylo (2001), 'Outcome Mapping: Building Learning and Reflection into Development Programs', International Development Research Centre, Ottawa.

EC (European Commission) (2013), 'Strengthening the Foundations of Smart Regulation – Improving Evaluation', COM(2013) 686 final, Brussels.

EC (European Commission) (2018a), 'Budget Support Guidelines', Tools and Methods Series Guidelines N°7, Brussels.

EC (European Commission) (2018b), 'Risk Management in the Commission: Implementation Guide (2018–2910)'.

EC (European Commission) (2019), 'Evaluation in Hard-to-Reach Areas', presentation summary, Brussels.

EC (European Commission) (2020a), 'EU Gender Action Plan (GAP) III – An Ambitious Agenda for Gender Equality and Women's Empowerment in EU External Action', SWD(2020) 284 final, Brussels.

EC (European Commission) (2020b), 'A Union of Equality: Gender Equality Strategy 2020–2025', COM(2020) 152 final, Brussels.

EC (European Commission) (2021a), 'Better Regulation Guidelines', SWD(2021) 305 final, Brussels.

EC (European Commission) (2021b), 'Circular Economy: Results and Indicators for Development'.

EC (European Commission) (2023), 'Better Regulation Toolbox', Brussels.

EC (European Commission) (2024), 'Evaluation with Gender as a Cross-cutting Dimension', Ares(2018)3264752, Brussels.

EEAS and EC (European External Action Service and European Commission) (2014), 'Evaluation Matters: The Evaluation Policy for European Union Development Co-operation', Luxembourg.

Fujita, Nobuko (2010), 'Beyond Logframe; Using Systems Concepts in Evaluation', Foundation for Advanced Studies on International Development, Tokyo.

Garred, Michelle, Charlotte Booth, Kiely Barnard-Webster and Ola Saleh (2018), 'Do No Harm and Gender: A Guidance Note', CDA, Cambridge, MA.

Gerring, J., and L. Cojocaru (2015), 'Case-Selection: A Diversity of Methods and Criteria', Monitoring and Evaluation News.

Goldwyn, R., and D. Chigas (2013), 'Monitoring and Evaluating Conflict Sensitivity: Methodological Challenges and Practical Solutions', UK Foreign, Commonwealth and Development Office, London.

Gough, D., S. Oliver and J. Thomas (2013), 'Learning from Research: Systematic Reviews for Informing Policy Decisions: A Quick Guide', Alliance for Useful Evidence, London.

Guijt, Irene (2014), 'Participatory Approaches', Methodological Briefs - Impact Evaluation No. 5, UNICEF.

Hassnain, Hur (2020), 'Mitigating the Risks of Remote Data Collection', Capacity4dev, EvalCrisis Blog #4, 20 July.

Hassnain, Hur, and Marco Lorenzoni (2020a), 'Disseminate till You Drop!', Capacity4dev, EvalCrisis Blog #9, 5 November.

Hassnain, Hur, and Marco Lorenzoni (2020b), 'Remote Data Collection for Evaluation and Research', Capacity4dev, EvalCrisis Blog #1, 9 June.

Hassnain, Hur (2021), 'Closing Learning and Feedback Gaps in Evaluation: How to Extend the Ownership of an Evaluation's Findings to Project Participants', *Evaluation Matters* 3.

Hassnain, Hur, Lauren Kelly and Simona Somma, eds. (2021), 'Evaluation in Contexts of Fragility, Conflict and Violence: Guidance from Global Evaluation Practitioners', IDEAS, Exeter, UK.

Haynes, L., B. Goldacre and D. Torgerson (2012), 'Test, Learn, Adapt: Developing Public Policy with Randomised Controlled Trials', Cabinet Office Behavioural Insights Team, London.

Hoogeveen, J., and U. Pape, eds. (2020), 'Fragility and Innovations in Data Collection', in J. Hoogeveen and U. Pape, eds., *Data Collection in Fragile States: Innovations from Africa and Beyond*, Palgrave Macmillan, London.

Humanitarian Outcomes (2021), Aid Worker Security Report 2021.

ICE (International Consulting Expertise) (2020), 'Study on Best Practices in Third Party Monitoring', European Commission, Brussels.

IDEAS (International Development Evaluation Association) (2013), 'IDEAS Code of Ethics'.

IMF (International Monetary Fund) (1999), 'Code of Good Practices on Transparency in Monetary and Financial Policies', Washington, DC.

International Alert (2004), 'Conflict-Sensitive Approaches to Development, Humanitarian Assistance and Peacebuilding: A Resource Pack'.

Jackson, E. T. (2013), 'Interrogating the Theory of Change: Evaluating Impact Investing Where It Matters Most, *Journal of Sustainable Finance and Investment* 3 (2): 95–110.

James, C. (2011), 'Theory of Change Review: A Report Commissioned by Comic Relief', London.

Kotu, V., and B. Deshpande (2014), 'Predictive Analytics and Data Mining: Concepts and Practice with RapidMiner'.

Marsh, David R., Dirk G. Schroeder, Kirk A. Dearden, Jerry Sternin and Monique Sternin (2004), 'The Power of Positive Deviance', *British Medical Journal*, 329 (7475): 1177–1179.

National Academy of Sciences (2012), *An Integrated Framework for Assessing the Value of Community-Based Prevention*, National Academies Press, Washington, DC.

OECD (Organisation for Economic Co-operation and Development) (2016), 'States of Fragility 2016: Understanding Violence', Paris.

OECD (Organisation for Economic Co-operation and Development) (2022), 'States of Fragility 2022', Paris.

OECD DAC (Organisation for Economic Co-operation and Development Development Assistance Committee) (2022), 'Protection of People Involved in Evaluation'.

OECD DAC (Organisation for Economic Co-operation and Development Development Assistance Committee) (2023), 'Glossary of Key Terms in Evaluation and Results Based Management', 2nd edition, Paris.

O'Neil, Cathy (2016), *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*, Crown, New York.

Patton, M. (2013), 'Utilization-Focused Evaluation (U-FE) Checklist'.

Patton, Michael Quinn, and Ricardo A. Millett (1998), 'Lessons Learned', The Evaluation Exchange: Emerging Strategies in Evaluating Child and Family Services IV (3/4): 14.

Pawson, Ray, and Nick Tilley (1997), *Realistic Evaluation*, SAGE Publications, London.

Scriven, M. (2008), 'A Summative Evaluation of RCT Methodology & An Alternative Approach to Causal Research', *Journal of Multi-Disciplinary Evaluation*, 5 (9): 11–24.

Shaxson, Louise, and Ben Clench (2011), 'Outcome Mapping and Social Frameworks: Tools for Designing, Delivering and Monitoring Policy via Distributed Partnerships', Working Paper 1, Delta Partnership, London.

Simister, N., and R. Smith, R. (2010), 'Monitoring and Evaluating Capacity Building: Is It Really That Difficult?', Praxis Paper 23, International NGO Training and Research Centre, Oxford.

Trevisan, M., and T. Walser (2014), *Evaluability Assessment: Improving Evaluation Quality and Use*, SAGE Publications, London.

UN ECOSOC (United Nations Economic and Social Council) (1997), 'Gender Mainstreaming', Extract from A/52/3, United Nations, New York.

UNEG (United Nations Evaluation Group) (2020), 'Ethical Guidelines for Evaluation', New York.

UNICEF (United Nations Children's Fund) (2011), 'How to Design and Manage Equity-Focused Evaluations', UNICEF Evaluation Office, New York.

UN Women (2020a), 'Good Practices in Gender-Responsive Evaluations', UN Women Independent Evaluation and Audit Services.

UN Women (2020b), 'UN Women Rapid Assessment Tool: To Evaluate Gender Equality and Women's Empowerment Results In Humanitarian Contexts', UN Women Independent Evaluation and Audit Services.

USAID (United States Agency for International Development) (2009), 'Glossary of Evaluation Terms', Washington, DC.

Van den Berg, Rob D., Penny Hawkins and Nicoletta Stame (2022), *Ethics for Evaluation: Beyond 'Doing No Harm' to 'Tackling Bad' and 'Doing Good'*, Routledge, New York.

Weiss, C. H. (1995), 'Nothing as Practical as Good Theory: Exploring Theory-Based Evaluation for Comprehensive Community Initiatives for Children and Families', in J. P. Connell, A. C. Kubisch, L. B. Schorr and C. H. Weiss, eds., *New Approaches to Evaluating Community Initiatives*, Aspen Institute, Washington, DC, pp. 65–92.

White, Howard, and Daniel Phillips (2012), 'Addressing attribution of cause and effect in small *n* impact evaluations: towards an integrated framework', 3ie, New Delhi.

Woodrow, Peter, Nick Oatley and Michelle Garred (2017), 'Faith Matters: A Guide for the Design, Monitoring & Evaluation of Inter-religious Action for Peacebuilding,' CDA Collaborative Learning Projects and Alliance for Peacebuilding.

Getting in touch with the EU

In person

All over the European Union there are hundreds of Europe Direct information centres. You can find the address of the centre nearest you at: https://europa.eu/european-union/contact_en

On the phone or by email

Europe Direct is a service that answers your questions about the European Union. You can contact this service:

- by freephone: 00 800 6 7 8 9 10 11 (certain operators may charge for these calls),
- at the following standard number: +32 22999696 or
- by email via: https://europa.eu/european-union/contact_en

Finding information about the EU

Online

Information about the European Union in all the official languages of the EU is available on the Europa website at: https://europa.eu/european-union/index_en

EU publications

You can download or order free and priced EU publications at: https://publications.europa.eu/en/publications. Multiple copies of free publications may be obtained by contacting Europe Direct or your local information centre (see https://europa.eu/european-union/contact_en).

EU law and related documents

For access to legal information from the EU, including all EU law since 1952 in all the official language versions, go to EUR-Lex at: http://eur-lex.europa.eu

Open data from the EU

The EU Open Data Portal (http://data.europa.eu/euodp/en) provides access to data sets from the EU. Data can be downloaded and reused for free, for both commercial and non-commercial purposes.

Publications Office
of the European Union