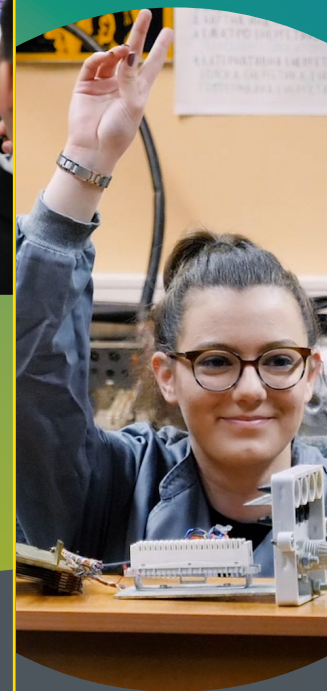


# GUIDELINES FOR IMPACT ASSESSMENT OF ACTIVE LABOUR MARKET POLICIES IN FRAGILE CONTEXTS





This report has been prepared by the European Training Foundation.

Author(s): Eva Jansova (ETF), Piotr Stronkowski (ETF)

Manuscript completed in December 2025.

When citing this report, please use the following wording:

European Training Foundation (2026), Guidelines for Impact Assessment of Active Labour Market Policies in Fragile Contexts, Publications Office of the European Union, Luxembourg

Luxembourg: Publications Office of the European Union, 2026

---

PDF: ISBN 978-92-9157-761-3 doi:10.2816/4048295

The contents of the report are the sole responsibility of the ETF and do not necessarily reflect the views of the EU institutions.

© European Training Foundation, 2026

Except otherwise noted, the reuse of this document is authorised under the Creative Commons Attribution 4.0 international (CC BY 4.0) licence (<https://creativecommons.org/licenses/by/4.0/>). This means that reuse is allowed provided appropriate credit is given and any changes are indicated. For any use or reproduction of photos or other material that is not owned by the European Training Foundation, permission must be sought directly from the copyright holders.



## PREFACE

As governments face budgetary restrictions and pressures to use public money in an effective and efficient way, there is a need to ensure that resources are spent on services that provide high economic and social return. Therefore, policies and programmes need to be monitored and evaluated in a regular and rigorous way, which allows for adjustments of existing policies and programmes to improve their performance vis-à-vis set objectives.

In order to support the consolidation of the thematic intelligence and exchange on active labour market policies (ALMPs) design, deployment and outcomes, the ETF initiative focused on the analysis of effectiveness of ALMPs in fragile context to draw specific recommendations for the design of ALMPs and for the improvement of monitoring and assessment of the implemented measures and services.

The key objective of this thematic paper is to present guidelines that provide practical and detailed instructions for the design and implementation of (impact) evaluations of

ALMPs. Taking into account the research cycle, it provides guidance on the different phases, such as (a) planning phase, including resources required (financial, human, research infrastructure), expected timeline, most common types of research design, and examples of data collection tools; (b) implementation phase; and (c) analytical and communication phase.

The guidelines were developed by ETF experts Eva Jansova and Piotr Stronkowski and are based on input gathered by means of several peer-learning events, including one titled “[Effectiveness of Active Labour Market Policies in Fragile Socio-Economic and Post-Conflict Contexts](#)”, held on 26-27 June 2024 in Brussels, and another titled “[Building Evidence Ecosystems in Education, Employment, and Training: A Peer Learning Event](#)”, which took place on 8-10 April 2025 in Cairo. The content also draws on background materials compiled for the thematic study “[Evaluating Active Labour Market Policies in Fragile Contexts: Insights and Lessons from Evaluation Studies in the ETF Partner Countries](#)”.

# CONTENTS

## 1

|                                    |          |
|------------------------------------|----------|
| <b>Introduction</b>                | <b>5</b> |
| 1.1 For whom                       | 6        |
| 1.2 For what                       | 6        |
| 1.3 How to navigate the guidelines | 8        |

## 2

|                                                   |          |
|---------------------------------------------------|----------|
| <b>Why evaluation in fragile contexts differs</b> | <b>9</b> |
|---------------------------------------------------|----------|

## 3

|                                                                               |           |
|-------------------------------------------------------------------------------|-----------|
| <b>Introduction to the impact evaluation of active labour market policies</b> | <b>12</b> |
| 3.1 Theory based impact evaluation                                            | 13        |
| 3.2 Causal attribution and (quasi-) experimental design                       | 16        |

## 4

|                             |           |
|-----------------------------|-----------|
| <b>Why evaluation fails</b> | <b>25</b> |
|-----------------------------|-----------|

## 5

|                                                                                        |           |
|----------------------------------------------------------------------------------------|-----------|
| <b>Planning evaluation</b>                                                             | <b>29</b> |
| 5.1 Evaluation plans as a tool for better integration evaluation with the policy cycle | 29        |
| 5.2 Defining scope of evaluation and evaluation objectives                             | 30        |
| 5.3 Selection of evaluation methods and tools                                          | 33        |
| 5.4 Setting timeline and budget                                                        | 38        |
| 5.5 Planning evaluation in the fragile context                                         | 40        |

## 6

|                                                               |           |
|---------------------------------------------------------------|-----------|
| <b>Implementation of evaluation</b>                           | <b>43</b> |
| 6.1 Commissioning evaluation and managing external evaluators | 43        |
| 6.2 Field work                                                | 48        |
| 6.3 Quality assurance for evaluation reports                  | 50        |

## 7

|                            |           |
|----------------------------|-----------|
| <b>Usage of evaluation</b> | <b>53</b> |
|----------------------------|-----------|

## 8

|                                                                         |           |
|-------------------------------------------------------------------------|-----------|
| <b>Strengthening institutions to properly design and use evaluation</b> | <b>58</b> |
|-------------------------------------------------------------------------|-----------|

## 9

|                   |           |
|-------------------|-----------|
| <b>References</b> | <b>62</b> |
|-------------------|-----------|





# 1

## INTRODUCTION

Evaluating active labour market policies (ALMPs) is essential to understand their effectiveness in improving employment outcomes and supporting social inclusion. As governments and international organisations invest significant resources into job training, employment subsidies or public work programmes, systematic evaluation should then ensure that such interventions achieve the intended goals.

Implementing evaluation in a rigorous way and using sound methodology is crucial in order to generate credible and reliable evidence. A well-designed evaluation ensures that the observed outcomes of interventions can be attributed to the measures themselves rather than external factors. This leads to well-informed decisions and an efficient use of public funds, while also strengthening accountability and trust among stakeholders. However, conducting evaluations

in fragile contexts may be challenging due to various factors, including the lack of existing data, feasibility of primary data collection, volatility of programme objectives or weakness of monitoring and evaluation systems. In such situations, the planned research should carefully address both the soundness of the evaluation methodology and the feasibility of its implementation.

These Guidelines are intended to provide practical instructions on how to design and implement impact evaluations of ALMPs, covering various stages in evaluation planning, implementation and usage. While most elements covered can be applied to evaluations of any type of programme or context, special attention will be given to conducting evaluations in fragile contexts and the specific considerations required in such situations.



### 1.1 For whom

The Guidelines are intended for key actors shaping and implementing active labour market policies, such as ministries, public employment services, but also international actors, such as donor organisations, or delegations to the European Union, which are often responsible for financing or implementing programmes aiming to support the employability of various target groups. Therefore, the guidelines do not aim necessarily to provide in-depth technical guidance but rather equip those who commission evaluations with sufficient knowledge to be able to procure and assess the quality of conducted research.

### 1.2 For what

The main objective of the guidelines is to provide practical guidance on how to design and implement evaluation of programmes supporting employment of different (vulnerable) groups, while paying specific attention to challenges in fragile contexts, such as poverty, social exclusion and instability due to conflict or weak governance, which may not be adequately addressed by standard evaluation approaches.

While there is a vast pool of resources providing (technical) guidance on impact evaluation methodologies both for researchers and practitioners, we attempt to contribute and build on this body of resources, addressing specifically the issue of evaluation of active labour market policies as well as the complexities brought by the fragility. Consequently, such guidelines aim at enhancing the capacity of actors responsible for commissioning such research and/or those operating in fragile contexts, thereby enabling them to plan, conduct or use evaluation results for designing interventions that are context-sensitive and impactful.



### 1.2.1 Box 1. Fragile context

In the field of development and global affairs, ‘fragile contexts’ refer to situations where states, systems, or communities face high risks and limited capacity to cope. Fragility can be defined as the combination of risk exposure and the inadequate ability to manage or mitigate said risks. It often leads to violence, poverty, inequality, displacement, and environmental or political decline.

#### The OECD (2022) measures fragility across six dimensions:

- **Economic:** Risks from weak economic structures, including macroeconomic shocks, youth unemployment, and poor fiscal management;
- **Environmental:** Exposure to climate change, natural disasters, pollution, and health crises;
- **Political:** Governance issues like lack of inclusiveness, transparency, and state legitimacy;
- **Security:** Threats from violence, crime, and weak law enforcement, including non-state armed groups;
- **Societal:** Social divisions from inequality, discrimination, and restricted civil society;
- **Human:** Barriers to health, education, and human development, particularly affecting vulnerable and marginalised groups.



### 1.3 How to navigate the Guidelines

These Guidelines are divided into several chapters. [Chapter 2](#) highlights specific challenges related to evaluation implementation, with a particular focus being placed on fragile contexts. [Chapter 3](#) introduces evaluation research, particularly impact evaluation design. [Chapter 4](#) addresses the reasons for which evaluations often fail, and emphasises the importance of strengthening institutional capacity for effective learning and evaluation. [Chapter 5](#) focuses on evaluation planning, including setting objectives, choosing the right methodology and budgeting.

[Chapter 6](#) delves into implementation, addressing, above all, how to commission evaluations, manage external evaluators, ensure quality and mitigate risks. [Chapter 7](#) looks at the importance of stakeholder engagement, effective communication of findings, and fostering a culture of learning from evaluation results and how lessons learned can be implemented going forward. In conclusion, [Chapter 8](#) addresses the role of institutions in the proper designing and usage of evaluations, particularly in fragile contexts.



2

## WHY EVALUATION IN FRAGILE CONTEXTS DIFFERS

Monitoring and evaluation of ALMPs in fragile situations is often considered challenging. The volatile and sometimes changing nature of these environments may lead to more dynamic programme objectives, as well as evolving monitoring and evaluation approaches. The labour market environment in fragile contexts is also likely to be complex, making it difficult to determine and measure the impact of interventions on ALMP outcomes.

In such settings, ALMPs are often contextualised and ‘sensitive’ so as to avoid unintended negative consequences. The monitoring and evaluation process must therefore not only evaluate programme outcomes but also assess how fragility-related risks affect programme delivery. In addition, general shifts in the country context must be monitored in order to determine whether the programme objectives, implementation modalities or partnerships remain appropriate and effective.

Challenges specific to the implementation of evaluation research in fragile contexts can be identified as follows (ETF, 2025):

### *Volatility, sensitivity and biases*

In fragile contexts, **conducting evaluations often requires adaptability as well as sensitivity to the local context**. This may include the need for changes in both the intervention logic as well as evaluation design, ensuring that the results remain timely and valid. Furthermore, evaluators must take into account potential sensitivities among all stakeholders involved, as certain findings or formulations can lead to tensions. Adhering to the ‘do no harm’ rule is essential, potentially requiring evaluators to exercise discretion or self-restraint. In addition, working in a fragile context may expose evaluators to stronger biases from the parties involved, which underscores the importance of methodological rigour and objectivity.





### *Limited time and resources*

In fragile contexts, impact assessment faces constraints in terms of limited time and resources across design, fieldwork and analysis phases. **Insufficient funding and contingency planning may undermine the ability to conduct robust evaluations** when unforeseen challenges arise. In such cases, researchers often face the need to arrive at a compromise between quality and the soundness of their methodology and the feasibility of its implementation. In addition, the urgency in delivering quick interventions may also result in programme implementation before establishing a comprehensive evaluation framework. This misalignment between the programme design and evaluation may limit the ability to establish clear baselines or indicators. In such contexts, it may be necessary to focus on basic, reliable evaluation methods rather than overambitious designs that may fail.

### *Limitations during evaluation implementation*

**Implementing evaluations of ALMPs in fragile contexts presents multiple challenges that may compromise the accuracy, representativeness, and ethical soundness of the research.** For instance, the recruitment of evaluators to work in high-risk areas is often difficult, and ensuring their safety also increases operational costs. Additionally, evaluators may need to have access to specific communities and target groups, which can be logistically and politically complex, and further influence the possibility to gather reliable data.

Moreover, methodological challenges – generally relevant for any type of evaluation – such as selection bias, the ethical concerns of randomisation, and insufficient sample sizes, can undermine evaluation outcomes. Traditional impact evaluation techniques rely on establishing valid control groups, which can in turn be both ethically and practically unfeasible in fragile settings. Small-scale interventions often lack the statistical power needed in order to reach robust conclusions.



Programme implementation challenges, such as low enrolment, high dropout rates or difficulty engaging vulnerable groups, not to mention high attrition rates, further complicate evaluations. Such problems can lead to self-selection bias, requiring advanced statistical methods to mitigate these biases and preserve the validity of evaluation results. Cost-effective strategies, such as carrying out in-person follow-ups and incentivising participants, have proven effective in some cases.

#### *Fragile institutions and limited evidence culture*

In fragile settings, public administration and civic institutions may pose challenges to conducting impact evaluations and making them actionable. **Limited institutional capacity can hinder the design, implementation and use of monitoring and evaluation** for active labour market policies. This includes issues such as limited accessibility or quality of data necessary for

effective monitoring and evaluation. These challenges may be intensified further by a weak culture of evaluation or, more broadly, the limited use of evidence for decision making. Consequently, there is often a dependence on donor bodies for the establishment and maintenance of systems for providing ALMP assessment.

Moreover, **representatives of state institutions may contribute only minimally to evaluation efforts due to constraints such as limited time, insufficient resources and competing policy priorities**. In addition, the prevailing institutional culture may undervalue the role of evaluation in learning, leading to the underutilisation of findings in programme improvement and policy development.



# 3

## INTRODUCTION TO THE IMPACT EVALUATION OF ACTIVE LABOUR MARKET POLICIES (ALMP)

Impact evaluation aims to quantify the contribution of an intervention, such as ALMP, to any changes observed. For instance, if unemployed individuals participate in training provided by public employment services and subsequently find a job, the relevant impact evaluation will assess the degree to which the training contributed to their employment.

Another important aspect of impact evaluation is understanding why the change occurs. Explaining the mechanism behind the change and the subsequent results is crucial. Specific questions include:

- Has the intervention made a difference?
- For whom, in what situations, and in which ways has the intervention made a difference?
- How much of a difference has the intervention made?
- To what extent can specific impacts be attributed to the intervention?
- How has the intervention made a difference?

There are two fundamental approaches to impact evaluation. The first focuses on determining whether the intervention has made a difference, while the second emphasises understanding why and how the intervention has had an impact.

The first approach forms the basis for counterfactual impact evaluation (CIE), which typically generates quantitative data showing the difference in outcomes between treatment and control groups, attributable to the causal effects of the intervention. Although regarded as the 'gold standard' of impact evaluation, CIE has limitations, namely in terms of inadequate consideration of differences across various groups or contexts, unsuitability for complex interventions, limited generalisability, and a focus on short-term and easily measurable outcomes. Despite these weaknesses, it remains methodologically robust.

The second approach is known as theory-based impact evaluation (TBIE). This method seeks



a deeper understanding of the mechanisms driving change and producing impact. It helps those responsible for designing and implementing interventions understand why interventions produce both intended and unintended effects, whom they impact, and in what context. TBIE has also been criticised for producing narrative data as opposed to numerical data, which may be perceived as less rigorous and reliable, and more susceptible to anecdotal evidence. Nonetheless, a rigorous methodological approach to TBIE can provide insightful knowledge about the intervention.

Consider a simple intervention: vocational training that leads to employment. Counterfactual impact evaluation measures the relationship between training and employment, while the theory of change delves deeper in order to provide additional insights, such as determining the characteristics of those participating in the training, how they were recruited and the reasons why, how training topics were selected and adjusted according to their needs, the quality of training, its implementation, the skills covered, and the implementation of those skills at work.

In practice, these two approaches are complementary. TBIE is considered a critical element of any evaluation, while CIE is necessary in order to test assumptions, which is a fundamental aspect of TBIE.

### **3.1 Theory-based impact evaluation**

The term ‘theory’ must first be explained in order to better understand the TBIE approach. ‘Theory’ is understood to be a set of assumptions about a given intervention and how such an intervention produces effects. In practice, we often deal with a set of theories, given that policy-makers, implementers, participants and stakeholders may have their own theories.

The basis of TBIE is the theory of change, which consists of a logical model presenting the logic of intervention: the relationships between resources – actions, outputs, outcomes and impact – and the set of assumptions as to why actions produce outputs, and contribute to outcomes and impact (see also [Chapter 4](#)). The key element is made up of those assumptions, as they explain the theory. For instance, the logic model can be used to show that training leads to skills acquisition and, subsequently, to employment.





This simple model is based on a number of assumptions, however, i.e. we assume that the reason for unemployment is the lack of or inappropriate skills; that training provided will be aligned with actual labour market needs, and both the potential and interest of the unemployed person. It can also be assumed that training has been properly designed in terms of its content, the relationship between theoretical and practical components, its length, and that the quality of training has been sufficient. The theory of change also takes into account external influences, including labour market dynamics, which can impact the observed change and contribute to the overall outcome.

A well-constructed theory of change improves the clarity of interventions by providing a better explanation of the ultimate objective, intermediate results, actions and assumptions. It also enhances critical thinking by identifying potential inconsistencies, gaps or vague assumptions.

A good example of TBIE would be implementing contribution analysis as one of the key approaches to the theory-based evaluation. The key question for contribution

analysis is whether the intervention contributed to the observed change and to what extent. The first step of the evaluation is to determine if it can be reasonably assumed that the intervention contributed to the change observed. This can be further broken down into two questions, i.e. what are the direct and indirect impacts of the intervention on the results.

The next step is the **development of a theory of change**. The theory of change is sometimes explicitly described in the documentation, but the description is often vague and incomplete. Additionally, documentation usually includes only the logical chain of intervention, lacking information on assumptions, mechanisms of change, risks and external factors. Furthermore, any official documentation typically determines the official version of the narrative. Therefore, at this stage, a comprehensive theory of change should be developed, including views on theories of change formulated by other actors.

The following step consists of **gathering evidence on the theory of change**, such as monitoring and evaluation information regarding actual actions taken and results achieved, evidence on casual links between components



of the theory of change, as well as other factors influencing results. Collecting such evidence aims to assess the credibility of the theory of change and the possible contribution that the intervention may have.

Based on the evidence gathered, the evaluator may **assess the theory of change**, identifying which links therein are strongly supported by evidence and which are weak, which may indicate little evidence available, weak logic or little agreement among stakeholders. At this stage, it is also important to identify differences in opinions among stakeholders regarding the theory of change, particularly linkages between actions and results.

Evaluation may involve **collecting additional evidence**, especially where the credibility of collected information is too weak to formulate a reliable assessment. This may include interviews, surveys, additional analysis, and the identification of existing evaluations and analyses. Based on these, the **intervention story should be revised** and, where possible, the assessment strengthened (whether positive or negative). The narrative that emerges should also include a credible explanation of the contribution of external factors.

As a result, the **intervention may be assessed as contributing to the observed change**

when it is built on well-founded and agreed-upon assumptions, supported by evidence and stakeholder consensus. Activities were implemented as planned, and the expected chain of results occurred and is supported by evidence observed, thereby confirming the theory of change. Other influencing factors should be considered and either ruled out as significant contributors or their role should be appropriately acknowledged. For a detailed summary of the contribution analysis, see the [better evaluation website](#).

Conducting **TBIE in fragile contexts** may be challenging, as these environments are often **complex** and **are liable to change quickly**. Multiple layers of interaction may not be apparent to external observers, while numerous factors can influence interventions and their outcomes. In these environments, standard **linear approaches to the theory of change may not be adequate**, and adopting a more dynamic method – one that recognises both intended and unintended effects, as well as a wide range of influencing factors – may be necessary. **Participatory approaches that do not rely on a predefined theory of change**



**are recommended in such situations.**

According to Better Evaluation, methods such as outcome harvesting, highlighting the most significant change and/or process tracing may be more suitable for evaluation in these contexts.

There are numerous advantages to TBIE, i.e. it enables a more comprehensive understanding of the perspectives and perceptions of different stakeholders, including those whose views might be overlooked in more standardised approaches. It also allows for a deeper insight into the mechanisms of change, explaining why an intervention has led to certain outcomes. Additionally, it helps identify other, at times unintended, outcomes and alternative methods of achieving the desired results. Importantly, TBIE facilitates a better understanding of how an intervention operates within a context specific to it, and the interactions between contextual conditions and the intervention itself. This approach is particularly suitable for complex interventions, which require deliberate and rigorous evaluation methods. However, TBIE does not provide a method for estimating the impact of an intervention or presenting it quantitatively. For this reason, counterfactual impact evaluation is required.

### **3.2 Causal attribution and (quasi-)experimental design**

In contrast to TBIE, the **CIE seeks to describe not only the changes resulting from the intervention, but also to determine if these changes can be attributed to the intervention itself** by examining causal attribution (or inference). To achieve this, we aim to compare the outcomes for a ‘treated group’ who have participated in a programme with a ‘comparison group’ of people, a control group, with similar characteristics who were not involved in the programme. The intention is to ascertain what would have been the results for beneficiaries had they not participated in a programme. A control group helps to assess how successful a newly tested programme has been by enabling the measurement of outcomes for programme participants when compared to non-participants in similar situations.

**In general, such a (counterfactual) impact evaluation would implement either an experimental or a quasi-experimental design.** While the experimental design relies on randomisation when it comes to the allocation of participants into either treatment or comparison groups, in the quasi-



experimental design approach, a comparison group is constructed by statistical means so as to control for differences between participating and non-participating individuals.

### *Experimental design*

#### **The random assignment of individuals into either treatment or control groups is the core of the experimental design.**

We aim to have two groups which are as similar as possible, where the only expected difference is the outcome being studied. Within experimental design, the randomised controlled trial (RCT) is a specific strategy in which eligible individuals are randomly assigned an intervention<sup>1</sup> or not, ensuring statistically comparable groups in terms of observed and unobserved factors. Data is then typically collected prior to, during and after the intervention in line with the predefined objectives and the theory of change of such interventions.

RCT, when run well, has the advantage of eliminating population bias. This, in turn, ensures that treatment and control groups are not systematically different from each other (considering both known and unknown factors) through random assignment. The results from an RCT can be analysed using standard statistical tools, with clearly defined participating populations.

#### *Counterfactual impact assessment through RCT*

If counterfactual impact evaluation and the implementation of RCT are feasible, several elements must be considered when planning for such an evaluation:

- *Randomisation strategy*, which ensures that the treatment and control groups are statistically similar while mitigating the potential selection bias. A few common approaches would be:
  - **simple random assignment**, where each eligible unit (individual, household, etc.)

<sup>1</sup>We may also opt for using more than one treatment group, and to have, for example, two groups, each receiving a different type of treatment.





has an equal chance of being assigned to either the treatment or control group. This type of randomisation can be described as flipping a coin for each unit;

- **stratified random assignment** where randomisation happens within specific strata based on sex, age or location, for instance. This increases the balance of groups according to key characteristics and may lead to more precise estimates;
- **cluster randomisation** where groups or larger units are randomised (e.g. schools, districts), which may be suitable in case individual randomisation is not feasible;
- **matched pair randomisation** where units are matched in pairs considering baseline similarities, and where one unit in each pair is randomly assigned to the treatment group.

Data collection process happens in different stages. Firstly, baseline data is collected, including typically socio-economic and demographic characteristics, employment history, household-related data and similar. Secondly, further data collection takes place during the intervention to monitor the adherence to the agreed protocol, identify implementation challenges or to check for attrition or spillover. Thirdly, post-intervention

data collection takes place at the end of the intervention to check for short-term employment outcomes, changes in behaviour or skills, and/or self-reported satisfaction, for example. Lastly, data on long-term outcomes, including information on labour market status, job quality, income levels, and overall well-being, may be collected months or years after the intervention.

In this case, we primarily rely on quantitative research methods, collecting quantifiable data through surveys or administrative sources that can be statistically analysed. However, incorporating qualitative research methods could provide additional context so as to help interpret the numerical results.

### Example – Wage subsidy

1. *Unit of assignment* – individuals registered at a public employment agency and qualifying for the wage subsidy;
2. *Assignment of individuals into treatment and control groups* prior to the start of the intervention, using stratified random assignment accounting for the strata related to sex and level of education;
3. *Collection of baseline data from both*



*treatment and control groups* prior to the start of the intervention, such as socio-economic situation, socio-demographic characteristics, household situation, etc., which may play an important role when evaluating the outcomes of wage subsidy intervention;

4. *Collection of data during the intervention* through a quantitative survey focusing on the evidence about the process. This is also the moment to check whether the random assignment holds, what is the level of dropout etc.;
5. *Collection of data on the outcomes of intervention* after having benefitted from

the wage subsidy measure through a quantitative survey, collecting data on the skills acquired, level of self-perceived confidence, job-search behaviours and other similar aspects;

6. A survey is administered sometime after the intervention to assess long-term outcomes. Data on outcomes related to labour market status, earnings, etc., is collected. Consequently, data and outcome indicators are compared across both groups by means of a comparison of averages, statistical testing to assess significance, or regression models to control for covariates.

Several preconditions must be met before implementing RCTs. **The key element is the need for large sample sizes to be able to draw valid results.** Both treatment and control groups require a minimum sample size of survey and administrative data. Card et al. (2011) provide guidance on the sample sizes required to measure a range of impacts. They draw a table in which each row represents the

employment rate of the comparison group, and each column represents the difference between treatment and control groups. For example, if the employment rate is 50% in the comparison group, the required sample size would be of 6 354 participants and the same number of non-participants in order to detect a significant impact of 2.5% in employment.



Besides collecting relevant data via surveys, administrative data sources may be used, as long as data protection provisions are met. Administrative data refers to information collected when people interact with public services, such as public employment services, health systems, social security agencies, educational institutes, taxation departments, and births, marriages and deaths registration offices. Related administrative data allows for the creation of large data sets and has become a vital tool for central and local governments conducting research (Smith et al., 2004). Administrative data can also be reused, it does not impose extra data collection burdens, and it allows for the use of larger population sizes than experimental and observational data. However, there are caveats to using administrative data. While open data allows transparency, participation, efficiency and economic innovation (Kitchin, 2014), the data collected is not always freely accessible and may be restricted to certain users (ADLS, 2016). National statistical data may be unavailable, incomplete or of low quality, particularly in fragile environments.

**In addition, RCTs need to be planned before an intervention begins and must be integrated into the implementation of ALMPs intended for evaluation.**

There are several potential challenges in the implementation of RCTs, especially in fragile settings. Planning must carefully consider time and cost, and special attention must be paid in order to ensure that the population studied remains representative.

Two major methodological challenges in these contexts include ensuring genuine randomisation and avoiding situations in which two different populations are compared, for example, when the treatment group has a higher average level of education than the control group because the assignment to the treatment group was based on a motivation to join an evaluated active labour market measure.

In addition, difficulties may also be connected with a low take-up of treatment, a change in the design of ALMPs during the evaluation process, challenges in tracing the researched units or a possible contamination of the control group due to other interventions, which in turn,





affect outcomes. Therefore, the feasibility of RCTs needs to be carefully considered and if preconditions are not met, other research strategies may be prioritised.

Lastly, randomised controlled trials (RCTs) and the process of randomly **assigning participants to treatment or control groups pose important ethical considerations**, as researchers determine which individuals receive an intervention – even among those who are eligible and in need – and those who do not.

#### *Quasi-experimental design*

When experimental design is not feasible, **quasi-experimental design can serve as a useful alternative by constructing a control group through statistical techniques**. Therefore, while such a design still allows for the study of the impact of selected intervention, random assignment does not feature. Beneficiaries take part in the intervention based on their selection or, for example, the assignment by PES staff, while the control group is created using statistical techniques aiming to be as similar as

possible, taking into account pre-intervention characteristics. The selected quasi-experimental methods are presented in **Box 1**.

This design is particularly pertinent when randomisation is not feasible, such as when evaluating the impact of an intervention retrospectively.

**The challenges or concerns associated with quasi-experimental design relate to the need for high-quality data and a sufficiently large sample size to construct a credible control group and accurately analyse the intervention outcomes.** As in the case for experimental design, it is vital to use administrative data, if available. Quasi-experimental designs, like all statistical approaches, rely on strong assumptions. In particular, evaluators must carefully select the variables used for matching treated and non-treated groups, which may mean that some unobserved confounding variables may be omitted, potentially biasing the results. Another key challenge is the selection bias, which arises when those who are ‘treated’ differ systematically from the ‘non-treated’ ones in ways that are not fully accounted for by the matching variables.





### **Box 1: Key quasi-experimental methods include:**

*Before and After Comparison of Programme Participants:* this approach measures changes in participants before and after their involvement in the programme, but it lacks a non-participating group for comparison, limiting the ability to account for external influences. Consider, for example, an evaluation of vocational training for the long-term unemployed where the initial situation of the group before the training – having been out of work for 12 months – and their labour market status 6 months after the training are compared. It may be the case that some of the beneficiaries become employed after gaining new skills and their respective situation improving. However, such an improvement cannot be attributed with certainty to vocational training, since other factors may have played a role, such as, for example, a general improvement in the economic situation of a region or country.

*Comparing participants' and non-participants' status after programme participation:* this method assumes that participants and non-participants are identical except for programme participation. Comparable groups are created based on observable characteristics (e.g. age, gender, education, employment history). However, limitations arise due to unobservable factors such as motivation or socio-cultural factors, which can introduce bias. Some evaluations supplement this with qualitative research to better understand variable meanings and reduce potential biases. Consider again the evaluation of vocational training provided to the long-term unemployed, where we analyse the labour market status of the long-term unemployed who took part in such a programme and those who did not. While the two groups may seem to be similar in terms of gender, education or age structure, other factors may influence the result, such as motivation to integrate the labour market as mentioned above, which



may explain the observed differences rather than the programme itself.

*Difference-in-differences approach (before and after comparison of participants and non-participants):* this measures the change over time of outcomes of participants as compared to non-participants. In this approach, it is assumed that if the programme did not exist, the two groups would have had identical trajectories over this period. One key challenge is that, for a number of reasons, this assumption does not hold. Revisiting the example of vocational training for the long-term unemployed, our research design would comprise of measuring those who take part in the training and those who do not before and after the intervention. Similarly to the case outlined above, there may be different factors that were not observed which may explain the measured results and not the intervention itself.

*Propensity score-matching:* individuals in the treatment group (participants) are statistically matched with non-participants who have similar observable characteristics (e.g. age, gender, education, employment history). The aim is to have two groups, with highly similar characteristics, so that the average difference in outcomes between matched individuals can be interpreted as the effect of the programme. This approach assumes that, depending on the observable variables used for matching, there are no remaining systematic differences between participants and non-participants that influence the outcome. This may be difficult to justify, particularly in fragile contexts where many unmeasured, context-specific factors (e.g. reliance on informal networks, insecurity) may influence the outcomes observed. Returning again to the vocational training example, if good administrative data exists, we can statistically match those who participated in training with those who did not, based on shared (observed)



characteristics. In simple terms, for instance, two women of the same age, level of education and years of work experience may be compared, where the only difference is that one has attended training and the other has not.

*Regression discontinuity design:* individuals are ranked based on specific measurable variables, usually socio-economic background variables. It is assumed that, after controlling for the ranking criteria, any remaining differences between individuals directly below or above the cut-off score are not statistically significant, and thus, the treatment effect can be estimated in the absence of significant bias. Revisiting yet again our example of vocational training

for the long-term unemployed, imagine that eligibility for the programme depends on a composite score based on measurable factors such as time spent unemployed, number of job interviews during the past six months, and similar criteria. A cut-off score is set; those scoring above it are admitted to training, while those just below are not. By comparing the outcomes in terms of labour market status (sometime after the intervention takes place) of those just above and just below the cut-off score, we can estimate the effect of the training programme for those close to the eligibility threshold.

Source: ILO 2018

#### Further reading

BetterEvaluation. BetterEvaluation knowledge platform. <https://www.betterevaluation.org/>

European Commission, Directorate-General for International Partnerships (2024), Evaluation handbook, Publications Office of the European Union. Available at: <https://data.europa.eu/doi/10.2841/340793>

European Commission (2013), Evalsed Sourcebook: Methods and Techniques, European Commission. Accessible at: [https://ec.europa.eu/regional\\_policy/sources/evaluation/guide/evaluation\\_sourcebook.pdf#page=172](https://ec.europa.eu/regional_policy/sources/evaluation/guide/evaluation_sourcebook.pdf#page=172)

White, H., Sabarwal S. & T. de Hoop, (2014), Randomized Controlled Trials (RCTs), Methodological Briefs: Impact Evaluation 7, UNICEF Office of Research, Florence. Available at: [Randomized\\_Controlled\\_Trials\\_ENG.pdf](#).

White, H., & S. Sabarwal (2014), Quasi-experimental Design and Methods, Methodological Briefs: Impact Evaluation 8, UNICEF Office of Research, Florence. Available at: [Quasi-Experimental\\_Design\\_and\\_Methods\\_ENG.pdf](#).



# 4

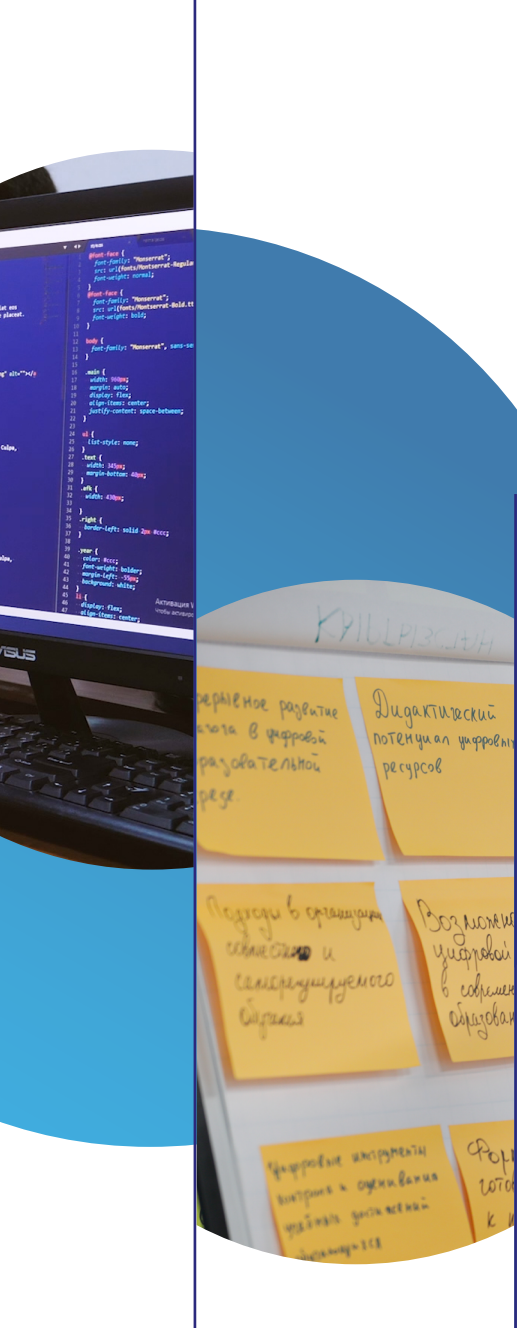
## WHY EVALUATION FAILS

It is well known that evaluation is an important element of the policy cycle, and it can contribute to better policy design, improved effectiveness and efficiency of public funds. However, despite significant investment in evaluation, it sometimes fails to provide sound results and actionable recommendations. It is difficult to estimate how often a given evaluation fails, but some evidence and the anecdotal experiences of authors suggest that can occur relatively often. The reasons for evaluation failure are manifold, i.e. from relating to the evaluation itself (its design, implementation, methodological soundness), to the knock-on effects of the institutional environment or political situation, etc..

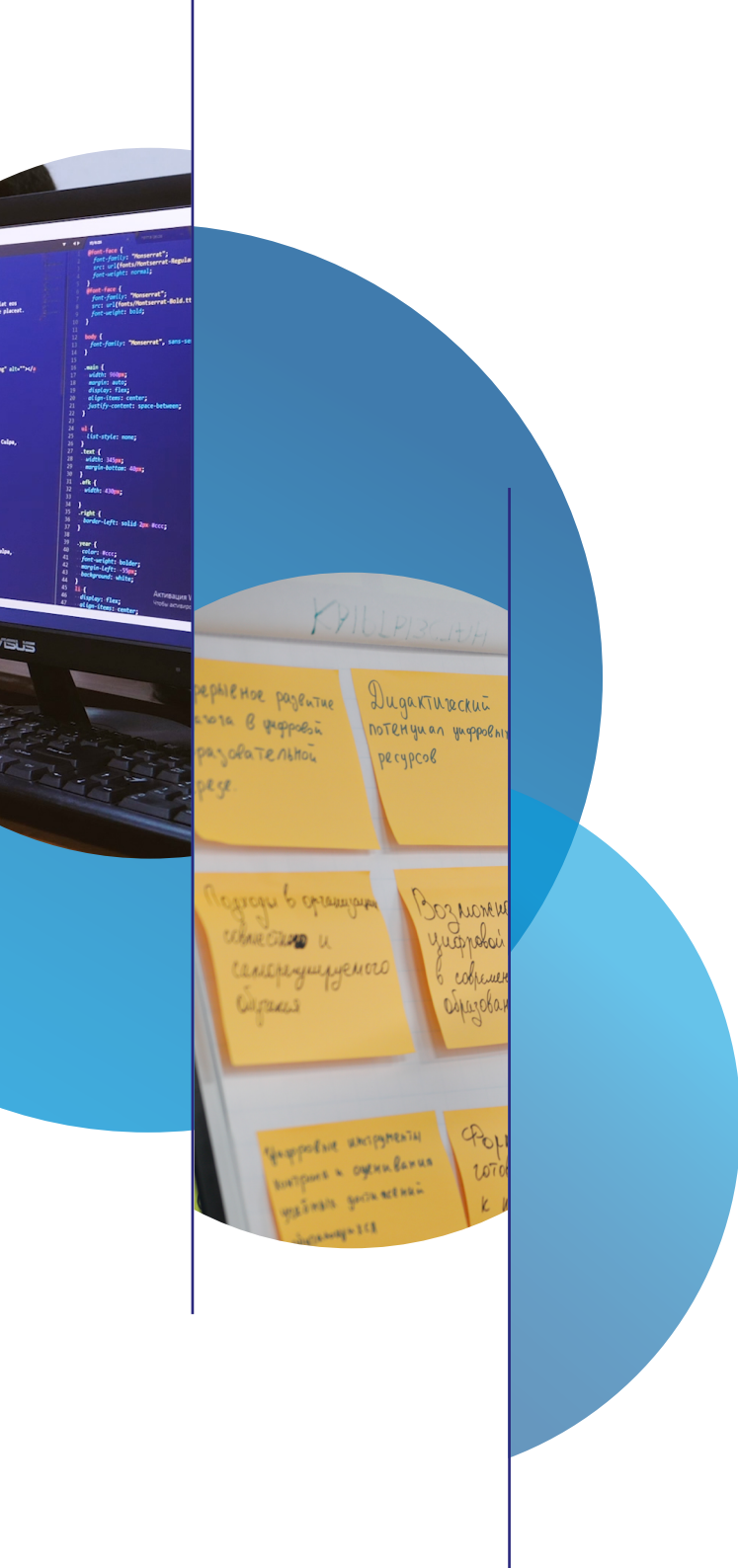
**Reflecting on the causes of evaluation failures is crucial, primarily with regard to gaining knowledge and experience, drawing conclusions, and implementing insights in the future.** Measuring how often evaluations fail to deliver expected outcomes can be challenging. Admitting that an evaluation has

been unsuccessful is also uncommon, and the reasons behind such failures are often complex. Given that many stakeholders typically participate, their perspectives on accountability may vary significantly. Furthermore, the success or failure of an evaluation can be considered from different perspectives, such as the implementation of the evaluation project, methodological relevance and rigidity, relevance and feasibility of evaluation findings, and their impact on policies or programmes. Nonetheless, evaluation failures offer substantial opportunities for learning (Hutchinson, 2019).

Anyone involved in evaluations knows that several issues may arise. Successful outcomes often depend on the experience and adaptability of evaluation teams from both the contracting and contractor sides in terms of anticipating problems and addressing them promptly.







### *Institutional capacity*

#### **Institutions may not yet be prepared to design, commission and use evaluations effectively for improving their strategic planning and programme implementation.**

This may be due to deficits within the organisation itself, such as a lack of openness to learning and using evaluation as a tool for improvement. The root of the problems may pertain to almost every aspect of evaluation, such as:

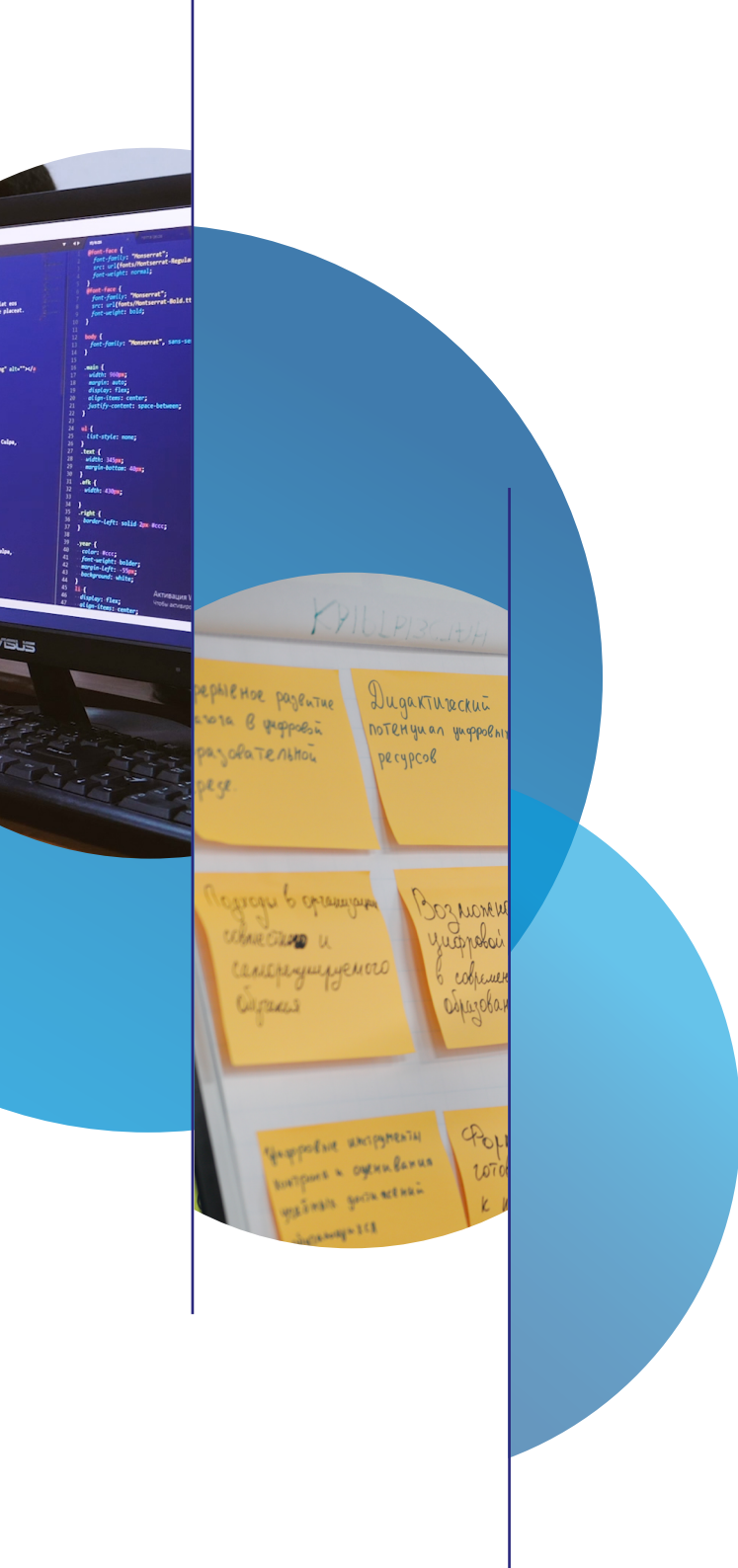
- Lack of skills for incorporating evaluation into strategic planning and implementation;
- Lack of skills for designing and implementing evaluation (such as knowledge about evaluation approach, methodology of social research, etc.);
- Underdeveloped culture of learning and improvement;
- The evaluation process may become ritualised, occurring solely to satisfy bureaucratic requirements, which are sometimes mandated by external donors;
- Perception of evaluation solely as a monitoring tool for legitimising the function and accountability of the organisation;

- Lack of sufficient human and financial resources, structure, and processes.

### *Evaluation is disconnected from strategic and organisational-level decision making*

**Evaluation practice is loosely aligned to strategic planning and thinking in an organisation.** In practice, although leaders of many organisations may declare the attachment the evaluation as a value, many problems may arise, including:

- Focusing on implementation questions, related to programmes or initiatives, instead of strategic questions (evaluation design);
- The needs of leaders and managers are not properly reflected in evaluation questions and data collected;
- The scope of evaluation is designed ad hoc, mostly influenced by individuals or units instead reflecting the whole organisation perspective;
- Lack of clarity on evaluation budget;
- Limited mechanisms supporting learning from evaluations, including capturing, storing, accessing and sharing learning lessons, conclusions and using them for improving practices (Preskill, 2013).



### *Problems on the evaluation side*

- Absence of a comprehensive evaluation plan

**A well-structured evaluation plan is crucial for guiding the evaluation process.** Without it, evaluations may lack direction, leading to inefficiencies and missed objectives. An effective plan outlines the purpose, methods, timelines, and responsibilities, ensuring that evaluations are systematic and aligned with organisational goals.

- Unclear or misaligned objectives and questions

**When evaluation objectives and questions are vague or not aligned with organisational goals, the evaluation may fail to provide meaningful insights.** Clear, specific, and relevant objectives are essential for guiding the evaluation and ensuring that findings are actionable.

- Inappropriate methods and tools

**Selecting unsuitable evaluation methods and tools can compromise the validity and reliability of findings.** It is essential to

choose methodologies that are appropriate for the context and objectives of the evaluation. Misapplication of methods can lead to inaccurate conclusions and ineffective recommendations.

- Low-quality or irrelevant data

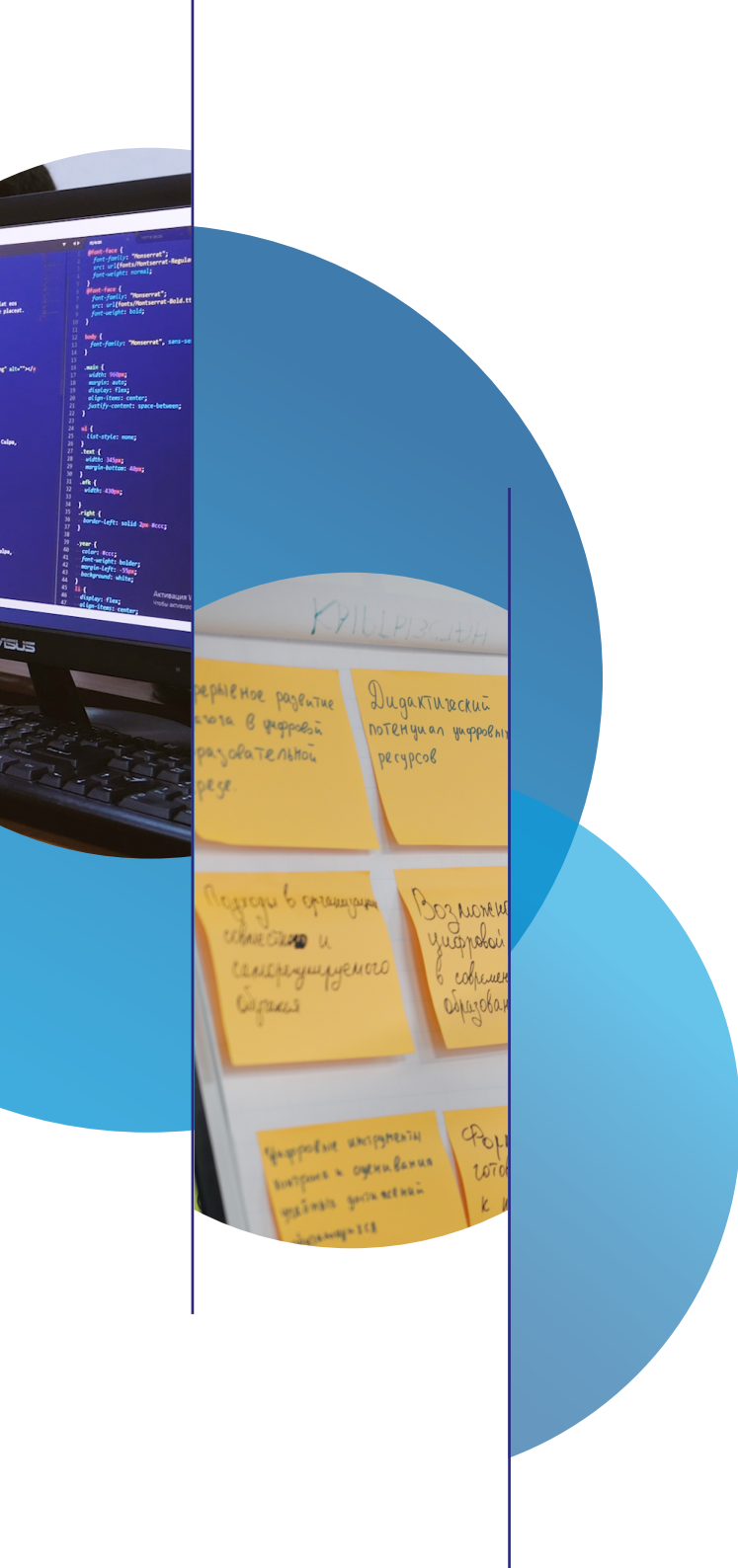
Data that is inaccurate, incomplete, or irrelevant can significantly hinder the evaluation process. **Poor data quality leads to unreliable findings and can misinform decision making.**

- Weak analysis and interpretation

Even with high-quality data, inadequate analysis can result in misleading conclusions. **Proper analytical techniques and critical interpretation are necessary to extract meaningful insights from data.** Failure in this area can render the evaluation ineffective.

- Ineffective presentation of findings

The way evaluation findings are presented can impact their utility. **Reports that are overly technical or too simplistic may not effectively communicate the results to**



**stakeholders.** Clear, concise and audience-appropriate presentations are essential for the practical application of evaluation findings.

#### ■ Resource constraints

**Limited financial, human, or time resources can impede the evaluation process.** Scarcity of resources may lead to compromised evaluation quality or incomplete assessments. Strategic planning and prioritisation are necessary to conduct effective evaluations within resource limitations.

---

#### Further reading

Hutchinson, K. (ed.) (2019), *Evaluation Failures: 22 Tales of Mistakes Made and Lessons Learned*, Thousand Oaks: SAGE Publications.  
Patton, M.Q. (ed.) (2012), 'Special issue: Stories of partial success', *Canadian Journal of Program Evaluation*, 27(3).  
Preskill, H. and Mack, K. (2019), *Building a strategic learning and evaluation system for your organization*. FSG. Available at: <https://www.fsg.org/resource/building-strategic-learning-and-evaluation-system-your-organization/>

# 5

## PLANNING EVALUATION

This chapter outlines important steps to consider when planning an impact evaluation of selected active labour market programmes. It specifically addresses the definition of evaluation objectives and research design. The chapter itself only covers the fundamentals; additional resources with more detailed information are provided at the end of this chapter.

### 5.1 Evaluation plans as a tool for better integration evaluation with the policy cycle

Planning an evaluation requires making decisions about the evaluation goals, key research questions or budget. These factors further influence decisions on research design, the scope of data to be collected, and the methods of data collection, as well as the type of analysis to be conducted on the data collected.

**Planning of the evaluation should take place before the implementation of programme or project begins, and all critical issues related to the evaluation should be considered.** It is advisable to prepare an evaluation plan, i.e. a structured document that

outlines how a programme, project or initiative will be assessed in order to determine its effectiveness, efficiency and impact. It serves as a roadmap for collecting, analysing and using data to inform decision-making for continuous improvement.

In general, the following questions need to be answered before the implementation of any evaluation: what do we want to find out? what should we measure? how? who will we collect data from? when will we collect data? or what will be done with the collected data?

Consequently, as for any research endeavour, the following elements should be outlined in the evaluation plan:

- a description of the programme/intervention and its context;
- evaluation governance (defining roles and responsibilities of different actors)
- Research design, including evaluation objectives, main research questions and hypothesis, methodology;
- sampling strategy (if relevant);







- data collection process, including data collection instruments (specifying how evaluation will be implemented and quality assured);
- data analysis process;
- dissemination plan and implementation of findings;
- timing and budget.

While an evaluation research design may employ various quantitative or qualitative strategies (see below for further details), it is critically important to determine in advance which data will be collected both before and during the intervention. This is especially crucial in the case of a randomised controlled trial (RCT), where the selection of control and treatment groups must occur prior to the intervention implementation. In addition, the evaluation plan should be developed in cooperation with the key stakeholders in order to facilitate ownership and subsequent usage of the obtained results.

In the next sections, we take a closer look at different elements that need to be developed and integrated in the plan, starting with main objectives and data collection methods, as well as practical implications related to the timeline and budget.

## 5.2 Defining scope of evaluation and evaluation objectives

Understanding the objectives of a programme is the starting point of the evaluation research, and helps us to define key research questions and further guide fieldwork.

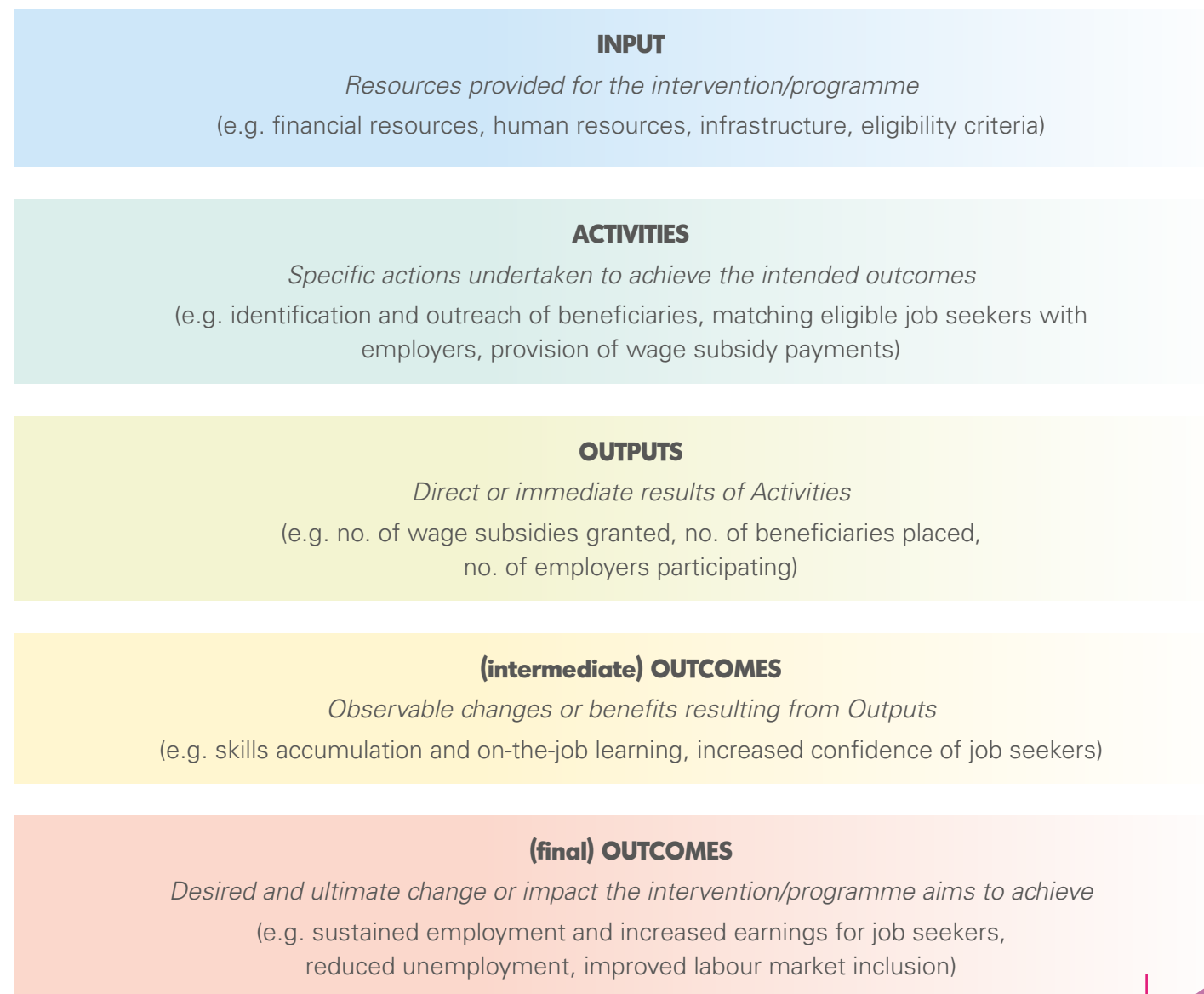
### ■ Clarifying programme objectives and goals – what do we want to know

The first step in the evaluation planning is to clarify the objectives and goals of the programme. **It is important to ensure that all stakeholders involved have a clear and shared understanding of what the programme aims to achieve and how it will do so.** Such clarity will help identify which components of the programme need to be evaluated.

At this point, it is crucial to clearly understand the logic behind the programme and build the *theory of change*, describing the causal link between the intervention/programme and the outcomes of such interventions. The key elements of the theory of change are described in [Figure 1](#).



**Figure 1: Theory of change – case of wage subsidy**





### ■ Developing evaluation questions

Evaluation questions are key and must be developed at the beginning of the research. Evaluation questions typically, but not exclusively, focus on four key areas:

- planning and implementation (e.g. how effectively was the ALMP planned and executed?);
- achievement of ALMP objectives (e.g., to what extent has the ALMP achieved its goals?);
- impact on beneficiaries (e.g., what difference has the ALMP made for participants?);
- broader impact (e.g., how has the ALMP affected the wider environment?).

Such questions are often organised under three main dimensions which guide the evaluation:

- appropriateness: does the programme address the identified need?
- effectiveness: is the programme achieving the intended outcomes?
- efficiency: are resources spent in a cost-effective way?

While the evaluation can comprise a wide range of questions, it is advisable to carefully select the most essential ones. A possible pitfall is the tendency to include too many questions or to concentrate on implementation details at the expense of examining actual impact. Lastly, as for the objectives, which are also key aspects of planned evaluations, research questions, in particular, should be consulted and defined together with the relevant stakeholders.



### 5.3 Selection of evaluation methods and tools

**The selection of research methods and tools is of course dependent on the key questions/objectives of the research and the type of causality we aim to analyse.**

At this point it is important to decide what data is to be collected to respond to the main research question, and to determine the most appropriate way to collect this data. Typically, the choice is to be made whether to employ quantitative or qualitative-research based approaches or their combination. Different data

collection methods can be considered, such as quantitative surveys of beneficiaries, qualitative interviews, observations, focus groups, documentation review and others.

Qualitative research focuses on obtaining an in-depth understanding of the research problem, is exploratory in nature, and emphasises the subjective interpretations of the participants. In contrast, quantitative research aims to quantify data and generalise results across populations by using structured tools such as surveys and statistical models (see [Table 1](#)).





**Table 1: Comparison of qualitative and quantitative research**

| Qualitative research                                                                                                                 | Quantitative research                                                                                                    |
|--------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------|
| The aim is a complete, detailed description.                                                                                         | The aim is to classify features, count them, and construct statistical models in an attempt to explain what is observed. |
| The researcher may only know roughly in advance what they are looking for.                                                           | The researcher knows clearly in advance what they are looking for.                                                       |
| The design emerges as the study unfolds.                                                                                             | All aspects of the study are carefully designed before data is collected.                                                |
| The researcher is the data gathering instrument.                                                                                     | The researcher uses tools, such as questionnaires or equipment to collect numerical data.                                |
| The data is in the form of words, pictures or objects.                                                                               | The data comes in the form of numbers and statistics.                                                                    |
| The subjective – individual's interpretation of events is important, e.g. it uses participant observation, in-depth interviews, etc. | The objective is to seek precise measurement and analysis of target concepts, e.g. using surveys, questionnaires, etc.   |
| The qualitative data is richer, more time consuming, and less susceptible to generalisation.                                         | The quantitative data is more efficient, more suitable for testing hypotheses, but may lack in contextual details.       |

Source: Veltri, 2011



Typically, the following data collection methods would be considered for evaluations (see [Further reading for more information](#)).

### **Quantitative surveys/questionnaires**

A quantitative survey is a set of (closed and open) questions aiming to collect information from respondents about their views and attitudes, as well as factual information relevant for the evaluation being conducted. Such surveys can be done face to face, online or by telephone. The information collected can then be deemed quantifiable and, in turn, analysed.

### **(In-depth) interviews**

In general, (in-depth) interviews aim to 'understand informants on their own terms and how they make meaning of their own lives, experiences, and cognitive processes' (Brenner, 2009:357). They usually take the form of long, focused interactions between beneficiaries and interviewers, which allows people to express themselves freely but also enables the influencing of the direction of the interviewing process as a whole. The format of (in-depth) interviews can take different forms, ranging from free/casual conversations

to highly structured interviews. The use of interviews may be helpful, for example, when evaluating services provided by a specific institution and being interested in the views and experience of different actors, such as management, staff and intended clients.

### **(Participatory) observation**

Studying objects in their 'natural' settings can be done through observation. Logically, the degree of participation of a researcher may range from that of a detached observer to that of an active participant in the group or activities being studied. It can also be done in a 'covert' or 'overt' way and differ in terms of the level of standardisation. The use of observation can be appropriate when, for example, we aim to study the dynamics of selected institutions or decision-making processes (Vesely et al., 2007).

### **Focus groups**

Focus groups are group discussions on a particular topic or set of topics. Such groups, which are moderated by a researcher, tend to be small and result in the collection of views/data on a selected issue. A researcher must



pay special attention to the structure they want to introduce into a discussion as well as to the composition of a group (Vesely et al., 2007). Such discussions are particularly useful in combination with other data collection methods (e.g. questionnaires). It can help a researcher to get a preliminary view into a problem and, consequently, improve survey tools by means of adding missing questions, clarifying concepts used, etc. At the same time, it can be used as a space for reflecting on the results obtained at the end of the research process.

### Document analysis

Qualitative research can also make use of existing documents, such as written texts (reports, legislation, diaries, correspondence, etc.), but also visual records (photos, videos, etc.). Through systematic document analysis, we may evaluate policies, procedures or even improve other data collection tools, such as surveys, through suggesting missing questions/topics.

More innovative data sources can be further considered, such as the use of big data, geospatial tools or secondary analysis of existing registries with administrative data.

**While each strategy, be it quantitative or qualitative, has its advantages, combining both approaches is often most effective,** particularly in fragile contexts where the collection of primary data through quantitative surveys may be difficult, for example, due to security issues. In addition, qualitative research and findings may play an important role in defining policy questions, looking in detail at how a programme or trial was actually implemented, helping to determine appropriate outcome measures by looking at subjective outcomes, clarifying what counts as effective or successful, or identifying and exploring unintended consequences. Furthermore, qualitative insights can enrich quantitative studies by informing the design of data collection instruments, assessing the fairness of comparisons in experimental studies, or unpacking variation within aggregated data (Davies, 2000).

### *Collection of data in fragile contexts*

**Collecting primary data in the field can be challenging in fragile settings due to factors such as security risks, restricted mobility, political instability or difficulties in locating beneficiaries.** In these contexts, traditional



data collection methods may not be feasible. Effective monitoring and data collection in highly fragile environments often necessitate flexibility – **accepting ‘good enough’ data** – and the adoption of innovative solutions. Such data limitations must be openly acknowledged in evaluations and how issues related to missing data or sources have been addressed in the reporting process must also be clearly outlined.

**If possible, it is advisable to employ mixed-methods approaches** and utilise nuanced instruments to ensure that relevant and context-sensitive information that supports robust analysis is captured. It is also important to consider external factors – such as **political, military, social, and economic influences** – that may affect intervention outcomes beyond its immediate scope, especially in rapidly evolving environments.

**Implementing evaluations in fragile settings necessitates specialised methodologies that emphasise safety, ethical standards, and meaningful local engagement.**

Primary data collection should employ instruments that minimise risk – such as secure, anonymous surveys, field-tested for

reliability – and be customised to respect local sensitivities. In areas of elevated risk, it may be prudent to collaborate with external parties to conduct discrete, ‘unbranded’ data gathering. Where feasible and appropriate, recruiting local evaluators and involving community members can enhance the relevance and quality of findings. However, such actions must be weighed against ethical considerations regarding independence and protection. Interview protocol should avoid topics that are politically sensitive, facilitate off-the-record sharing, and incorporate confidentiality measures like the Chatham House Rule. Enumerators should possess contextual awareness and receive training in terms of how to navigate complex social environments without endangering themselves or participants. This may include involving beneficiaries in the monitoring process or utilising advanced techniques such as remote monitoring.

**Remote data collection via digital tools can broaden access in insecure contexts, though such approaches require careful consideration.** Technologies including phone surveys, GPS-enabled devices and online interfaces may yield cost efficiencies





and improved data quality, but also present challenges related to the exclusion of digitally marginalised groups and a potential overreliance on quantitative methods. Complications such as limited personal interaction, diminished trust, barriers to technical literacy and apprehensiveness about surveillance or data misuse must be addressed.

Emerging techniques, such as narrative-based tools (e.g. SPROCKLER, SenseMaker) and Earth observation data, offer additional avenues for capturing both human and environmental variables, particularly in remote or contested regions.

#### **Leveraging existing administrative data –**

such as tax records, education statistics, or civil registries – can provide valuable insights for evaluations, provided that these sources are available and reliable. Such types of data can be considered if quasi-experimental design is used (see above). Big data sources, including mobile phone metadata, satellite imagery, or social media analytics, may also be considered. While such approaches have limitations related to data quality, coverage, or privacy concerns, they can be valuable alternatives when

traditional data collection is no longer viable.

#### **Selecting appropriate methodologies and tools demands comprehensive contextual analysis.**

The ‘5 As’ framework – Availability, Affordability, Awareness, Ability and Agency – serves as a valuable guide for identifying suitable ICTs and recognising obstacles to broad participation.

### **5.4 Setting the timeline and budget**

#### **■ Establishing an evaluation timeline**

Evaluation should be included in the initial planning stages of an ALMP. This process involves making decisions on:

- Key evaluation questions (see above);
- Evaluation design (identifying data needed to answer the questions);
- Type of data collection (evaluation methods used for data collection).

In order to accurately assess the impact of a measure, sufficient time must pass for both the implementation of the intervention and for outcomes to materialise after the intervention has taken place. This is particularly important



for prospective or real-time evaluations, which is the case of RCTs. For example, when evaluating a vocational training programme, data collection should be delayed for at least 6 months from the end of the intervention. Ideally, outcomes should be measured at multiple time points (e.g. after 6, 12 and 24 months) to capture not only employment status but also the stability and quality of employment.

Taking a simple example of a vocational training measure for the unemployed, using an RCT design and survey data, we aim to measure outcomes such as labour market status and wages. The possible evaluation timeline could include:

- Months 1–3 → Preparation and agreement on an evaluation plan;
- Months 4–5 → Collection of baseline data & randomisation (if RCT);
- Month 6 → Intervention takes place (training, four weeks);
- 6 months after intervention (Months 12–13) → Post-intervention data collection;
- Months 14–16 → Data cleaning, analysis, and reporting.

### ■ Setting an evaluation budget

An evaluation budget is essential for internal programme budgeting and commissioning external evaluations. When creating a budget, consider:

- Staff salaries based on time spent on evaluation;
- Fees for external consultants;
- Data collection costs (e.g. materials, focus groups, survey software);
- Incentives for interview participants (e.g., vouchers);
- Production costs for outputs (e.g., reports, presentations, videos, infographics);
- Costs related to dissemination workshop(s);
- Costs related to capacity building designed to strengthen capacities of public institutions in planning and commissioning impact evaluation (if applicable).

Factors influencing the budget include organisational capacity, logistical issues, field visits, interviews (particularly face-to-face interviews) in remote or conflict-affected areas, and evaluation complexity. Sophisticated methods,



larger sample sizes, the number and type of deliverables, as well as longer research time frames, will increase costs.

### 5.5 Planning evaluation in the fragile context

Preparing impact evaluations in fragile contexts involves a unique set of challenges due to weak institutions, limited resources, and volatile environments. These conditions demand flexible, context-sensitive approaches to ensure meaningful evaluation and learning. Below are the key challenges and possible solutions:

*Thorough preparation:* carrying out evaluations in fragile contexts requires significantly more detailed planning and preparation to ensure that all aspects of fragility are fully considered.

- Conduct conflict analysis to examine background, history, and causes of conflict, as well as identifying relevant groups, power structures, and potential interactions between the evaluation process and the conflict context. Additionally, engaging in stakeholder analysis is advised so as to identify all parties interested in or affected by the evaluation, including policymakers, those implementing it, beneficiaries, and opposing groups;

- It is advisable to conduct a preparatory (scoping) mission to engage relevant stakeholders, clarify evaluation procedures, and, if needed, test data collection instruments. This process facilitates the identification and management of potential unintended outcomes, including those that may affect vulnerable stakeholders or evaluation team members.

*Capacity constraints:* in fragile contexts, local technical expertise may be limited, infrastructure may be weak, and human resources insufficient for rigorous monitoring and evaluation.

- Partner with local universities or international institutions to fill technical gaps.

*Weak institutional structures:* governments may lack ownership or the capacity to manage ALMP evaluations.

- Promote government ownership and involvement in evaluation planning and implementation. Offer training to policymakers and technical staff so as to build long-term institutional monitoring and evaluation capacity.



*Donor priorities in conflict with national ones:* donor priorities and evaluation frameworks may not align with the context of the host country.

- Foster ongoing dialogue between donors and those implementing the evaluation. Align evaluation design with both donor requirements and local policy goals.

*Insufficient budgets:* evaluation is often underfunded, thus limiting the scope and quality of impact assessments.

- Ensure evaluations are built into programme design and adequately budgeted. Use cost-effective methods like administrative data analysis where appropriate.

*Unclear evaluation parameters:* ambiguity in evaluation goals and methods leads to mismatched expectations and irrelevant findings. It is essential to consider the complexities of the fragile context when setting evaluation objectives and formulating research questions.

- Clearly define research questions, success criteria, and data requirements early in the process. Ensure evaluators understand the context on the ground and ensure the involvement of key stakeholders throughout the evaluation process;
- Evaluation questions should cover sources of tension, conflict drivers, key actors and peacebuilding opportunities, while acknowledging that external assistance could unintentionally contribute to tensions.

*Control group and ethical issues:* constructing control groups may be impractical or unethical in fragile settings.

- Use randomised micro-segmentation where possible or alternative comparison methods that include all participants. Prioritise inclusive and conflict-sensitive evaluation designs.





---

#### Further reading

Bakrania, S., Balvin, N., Daidone, S. and De Hoop, J. (2021), Impact evaluation in settings of fragility and humanitarian emergency, Innocenti Discussion Paper 2021-02, UNICEF Office of Research – Innocenti, Florence. Available at:

<https://doi.org/10.18356/25211110-2021-02>.

BetterEvaluation (2024), Monitoring and evaluation in fragile, conflict-affected and violent (FCV) settings. [online] Last updated 18 September. Available at:

<https://www.betterevaluation.org/methods-approaches/themes/monitoring-evaluation-fragile-conflict-affected-violent-settings>.

BetterEvaluation. BetterEvaluation knowledge platform.

<https://www.betterevaluation.org/methods-approaches/methods/questionnaires>.

<https://www.betterevaluation.org/methods-approaches/methods/interviews>.

<https://www.betterevaluation.org/methods-approaches/methods/focus-groups>.

<https://www.betterevaluation.org/methods-approaches/methods/participant-observation>.

European Commission (2013), Evalsed Sourcebook: Methods and Techniques, European Commission. Accessible at:

[https://ec.europa.eu/regional\\_policy/sources/evaluation/guide/evaluation\\_sourcebook.pdf#page=172](https://ec.europa.eu/regional_policy/sources/evaluation/guide/evaluation_sourcebook.pdf#page=172).

Hassnain, H., Kelly, L. and Somma, S. (eds.) (2021), Evaluation in Contexts of Fragility, Conflict and Violence: Guidance from Global Evaluation Practitioners, Exeter, UK: IDEAS. Available at:

<https://ideas-global.org/wp-content/uploads/2021/06/EvalFCV-Guide-web-A4-HR.pdf>.

# 6

## CARRYING OUT AN EVALUATION

### 6.1 Commissioning an evaluation and managing external evaluators

The evaluation can be conducted by an internal evaluation team, i.e. an evaluation unit. This is the case for the **Development Impact Group** as part of the World Bank. However, a more typical situation is that the evaluation is outsourced to an external organisation, company or group of experts. It is acknowledged that such an approach provides the best assurance for independent evaluation.

Commissioning an evaluation with external experts may present a number of advantages:

- **Ensures that evaluators possess adequate skills, knowledge, and experience** – this is particularly important if such resources are lacking in commissioning institutions;
- **Enhances external analysis and assessment of the intervention** – persons

involved in the design and implementation of the intervention may be biased and have difficulties providing an impartial assessment;

- **Brings knowledge about the subject of the evaluation, similar projects or interventions and their results, as well as available resources and networks.** External evaluators can provide a broader, fresher perspective on the intervention, which may enrich the evaluation process;
- **Boosts the usability of evaluation results** – evaluation experts usually have strong skills in communicating findings, balancing the necessary level of technical detail with clear conclusions and actionable recommendations for interventions.

However, one should be aware of the disadvantages and risks related to cooperation with external evaluators, for instance:





- **External evaluators usually lack insider knowledge** and may not understand all of the subtleties associated with planning and implementing an intervention;
- **Evaluators' knowledge about the evaluation topic may be shallow and insufficient** to provide an in-depth analysis of the intervention. Their knowledge and skills in evaluation methods may also be inadequate;
- **Staff may perceive external evaluators as sources of oversight or disruption**, which can lead to limited responses and make it difficult to obtain candid and detailed answers;
- **The transaction costs of managing external evaluators can be high.** If expectations, division of responsibilities, or quality criteria are unclear, problems may arise that hamper the process and limit the usefulness of the evaluation.

These considerations show that evaluations must be carefully designed and managed to maximise their value.

- **Establishes a strong evaluation leader within your organisation** – this can be a person or a team responsible for preparing

and managing the evaluation process, including maintaining relationships with the evaluators. Even in cases in which the evaluation is conducted by external evaluators, the internal team must be fully involved, and responsibility for the final outcome lies with both sides. It is essential to ensure that the internal evaluation leader has sufficient resources and capabilities:

- **The possess solid knowledge of the intervention and the broader field of intervention** (e.g. the labour market), as well as an understanding of the expectations and needs of future users of the evaluation;
- **They have knowledge and skills in social research methodology and research practice** to act as a partner for evaluators in discussions and to respond when evaluators demonstrate an insufficient level of expertise;
- **They have experience in managing projects, particularly research projects**, and the authority to make day-to-day decisions regarding the evaluation;
- **The act as a bridge between evaluation users and evaluators.** The person (or team) in this role must understand the



needs of key stakeholders to ensure that these are taken into account when commissioning the evaluation.

- **Ensures that evaluators have a clear understanding of their tasks** and have access to all necessary information and data – this is often a critical aspect of cooperation with external evaluators.
- **Evaluators must have a solid understanding of the theory of change of the intervention** and the specific aspects they are expected to evaluate: which changes or effects should be measured, how they should be measured, etc. This must be clearly communicated to them. However, one of the evaluators' tasks should still include analysing the theory of change, incorporating the perspectives of other stakeholders, not only those of the institutions responsible for the intervention;
- **Evaluators should be informed in advance about which data will be provided** by the contractor, and which information must be collected independently by the evaluation team. It is particularly important to anticipate and secure access to contact and personal

data if such information is required for carrying out the research;

- **Evaluators need clear information on the evaluation timeline, allocation of responsibilities, and quality standards for deliverables.** In some cases, attention to specific details is crucial – for example, determining who will invite participants to workshops. Evaluators may also need to adapt to changing circumstances, especially in fragile contexts. However, it is important to remember that they must operate within established budgets and time frames.

### 6.1.1 Selecting a team of evaluators

The selection of a team with appropriate experience, skills and knowledge significantly influences the quality, relevance, and usability of an evaluation. Even when an evaluation is well designed, has high-quality data, and utilises effective processes, contracting an unsuitable team may result in outcomes that do not meet expectations. Conversely, even if the evaluation concept and settings are less than ideal, a competent team can often optimise available resources to produce satisfactory results. This is particularly true in a fragile context, in which quick adaptation to changing environments may be necessary.





Most evaluation handbooks emphasise the need to form a team with an appropriate combination of skills and knowledge. Achieving a balance between policy understanding and technical expertise is important for effective impact evaluation. The following are several principles for forming an evaluation team:

- Include **a team leader** with substantial experience in managing evaluation research. This is particularly important when the evaluation involves significant fieldwork and randomised controlled trials (RCTs), which can be complex to manage and present various risks. Effective managerial skills are especially relevant in fragile environments, where adaptability is required due to frequent changes;
- Ensure **sector-specific knowledge** – at least one team member should possess a solid understanding of the subject under evaluation, such as labour market policies, vocational education and training (VET), adult education, etc. If necessary, engage experts with even more specialised knowledge. Involving a local expert with sector-specific insights may be beneficial, as they can contribute contextual understanding regarding the operation of systems and the feasibility of recommendations. Evaluators with extensive experience in the relevant field and context tend to deliver more detailed and deepen observations and suggestions;
- Maintain strong technical skills in impact evaluation – particularly for impact assessments, it is important that the team demonstrates a high level of competence in econometrics or statistics. Team members should have experience with **sampling, survey design, power calculations**, and data quality assurance;
- The team should include **field coordinators and field experts** with practical skills related to the implementation of evaluation methods on site. They should possess strong organisational abilities, especially for RCTs, along with local knowledge, language skills, and data collection experience. Their expertise should align with the evaluation design; if the evaluation has a significant qualitative component, it is important to involve individuals with extensive qualitative research experience;
- Ensure independence of the evaluation team from the implementing agency so as to avoid bias.



A few aspects need to be considered when contracting the evaluation team:

■ **Achieving a balance between local and international experts is recommended.**

Generally, local experts are prioritised if they have the necessary sector-specific, technical or managerial skills. Local experts often have greater familiarity with the local context and may operate more effectively within it. Engaging local experts can also support the development of local evaluation capacity and culture. International experts may be engaged when specific expertise not available locally is required;

- During the initial phase of implementation, early team coordination is advised: defining **roles, responsibilities, communication protocols, timelines** and quality standards. If needed, training for the team should be included to ensure all members are adequately prepared for the research. Training is particularly important for those conducting fieldwork to ensure full understanding of relevant concepts and to provide opportunities to address any questions about the research tools.

**The selection of the team should be based on clear and transparent criteria,** which

should be included in the terms of reference. The verification of the knowledge and skills is usually done on the basis of prior experience. Therefore, one must always require detailed information regarding prior experience with clear references to said requirements. Keep in mind that it is always the responsibility of the contractor to establish that experts meet requirements. If you have doubts, always ask the contractor for clarification. Another good practice would be to assess past performance in similar evaluations. You can review examples of past evaluations or contact references to assess quality, timeliness and collaboration.

**A transparent and competitive selection process is commonly used.** Procedures often include an open call for proposals or creating a shortlist based on prequalified providers. The assessment generally consists of two stages: reviewing the technical offer (including methodology, team expertise and approach) and evaluating the financial offer (such as cost-effectiveness).



The selection is usually based on multi-criteria assessment, in which varying weightings are assigned to specific criteria, i.e.:

- Technical expertise (40-60%);
- Relevant sector knowledge (10-20%);
- Team qualifications and experience (20-30%);
- Methodological soundness (20-30%);
- Cost and value for money (10-30%).

**Involving stakeholders in the selection process**, such as programme managers, implementing partners, or evaluation users, is a common approach. This can be done by forming a steering committee or an advisory group. Such involvement may facilitate ownership and increase the probability that results are utilised.

## 6.2 Fieldwork

Contracting agencies are not typically involved directly in fieldwork. Nevertheless, the quality and relevance of evaluation findings depend greatly on the experience, skills and thoroughness of evaluators, as well as the quality of fieldwork implementation. The improper use of evaluation methods and tools can significantly undermine the robustness

of data, thereby affecting findings and recommendations.

In fragile contexts, it is especially important to balance the affordability and reliability of collected data. While every measurement helps to reduce uncertainty, and while having imperfect data is often better than having none, it is crucial to recognise that inaccurate results can be damaging. False negatives may result in the reduction or cessation of interventions, while false positives could lead to overinvestment of limited resources in policies or programmes that are not truly effective.

In order to facilitate the smooth implementation of fieldwork, at least one member of the team should have sufficient experience in fieldwork management (see [section 6.1.1](#)). It is also vital to ensure that the contractor has adequate capacity to conduct high-quality fieldwork, for example, by means of a network of experienced enumerators.

Engaging local experts and enumerators is often an effective approach. They usually have a better understanding of the local context and are more adept at navigating complex or ambiguous situations. However, any and all



sensitive issues related to their involvement must be carefully considered. Careful attention should also be paid to the wording of questions in research tools to avoid sensitive topics. Enumerators must be trained in conflict sensitivity and gender responsiveness, and it is essential to secure informed consent from all respondents. Evaluations should focus on building trust, encouraging community participation, and supporting local capacity wherever possible. Data triangulation, inclusive validation processes, and robust feedback mechanisms are crucial for ensuring that results are both credible and actionable.

Enumerators and subject matter experts must be thoroughly prepared for the research process. The team should have a clear understanding of both the logical framework of the intervention and its study objectives, as well as the importance of data quality. Organising comprehensive training sessions for field staff, together with maintaining proper documentation and transparency through field manuals, enumerator guidelines, data dictionaries and coding protocols, is highly recommended.

The field team should be equipped to address ethical questions and maintain data privacy, which is particularly important in randomised controlled trial (RCT) designs where complex ethical dilemmas may arise. Training and guidance on data privacy – including informed consent, secure storage, and anonymisation procedures – are necessary for all field staff.

Any risks related to fieldwork should be identified during the inception phase and documented in the inception report. The risk assessment should cover risks associated with both data quality and contextual fragility. A thorough assessment of the fragile context, as outlined in Section 5.5, may be required. The risk identification process should also specify methods for mitigating any identified risks. However, it is important to remain mindful that in fragile contexts, circumstances may change rapidly and new, unforeseen risks may emerge. As such, a flexible approach to budgeting, timelines and implementation should be maintained. Risks related to staff involved in fieldwork should be minimised as far as possible.





When carrying out fieldwork, it is important to implement quality assurance mechanisms. Enumerators should be supervised by a team leader or supervisor. Regular spot-check interviews and reviews of completed forms are recommended. Debriefing sessions can help address issues, clarify questions and identify inconsistencies. Quality control may also involve re-contacting a sample of respondents to verify the accuracy of the data. When using computer-assisted personal, telephone or web interviews, tools such as automatic filters and logic checks (e.g. checking variable types or ensuring age entries fall between 0 and 120) can help reduce errors. These systems may also record the location, time and duration of interviews to detect potential irregularities, such as those conducted at unusual hours or in very short intervals.

To promote data accuracy, a phase of data cleaning and validation after fieldwork is advisable. This process usually includes checks for consistency, outliers, missing values, logical errors, and the detection of duplicate or fabricated data (such as repeated entries or suspicious patterns). Triangulation – comparing findings with alternative sources like administrative data – is also beneficial.

Additionally, statistical checks such as analysing distributions and comparing results against known population parameters should be carried out at this stage.

### 6.3 Quality assurance for evaluation reports

Effective quality assurance (QA) mechanisms are essential in achieving high standards and reliability throughout the evaluation research process and its outputs. Quality assurance should be integrated from the outset, beginning with the evaluation design phase. It is strongly advised that both the contracting agency and the evaluation team establish comprehensive QA mechanisms.

This process is considerably streamlined when an organisation has a pre-existing Evaluation Quality Assurance System (EQAS). For instance, the International Organization for Migration, a United Nations agency focused on migration challenges, offers a robust system. Their materials are available [here](#).

Notably, the system delineates key roles and responsibilities within the evaluation process and includes quality criteria in the form of checklists covering Terms of Reference,



inception reports and final reports. Although the framework is straightforward, it serves as an effective model for implementing procedures that support stakeholders in delivering high-quality evaluations.

A more elaborate framework is detailed in [The Aqua Book: guidance on producing quality analysis for government](#), developed by the Government of the United Kingdom. This approach asserts that quality assurance is an overarching process that must be embedded at every stage of analysis within public institutions. The guidance highlights two main requirements for effective quality assurance:

**Environment:** establishing a culture that prioritises quality, welcomes constructive feedback, provides clearly defined responsibilities, and allocates sufficient time for thorough review.

**Process:** developing structured procedures for each phase of analysis, involving close collaboration with commissioners and users to ensure a shared understanding of the analysis objectives and limitations.

A comprehensive QA system for impact evaluation typically encompasses the following elements:

- **Checklists:** clearly defined standards and step-by-step checklists spanning from evaluation design to the final report;
- **Standardisation:** standardised tools, templates (e.g. terms of reference, reports), and technical notes offering detailed instructions wherever feasible;
- **Roles and Responsibilities:** explicit definitions of team members' roles and responsibilities, particularly those related to quality assurance;
- **Stakeholder Involvement:** active engagement of stakeholders, especially intended users of the evaluation, from project design through inception and reporting, to ensure suitability for purpose;
- **Peer Review:** implementation of peer-review mechanisms at all stages of evaluation. Reviews can be conducted internally by team members or, ideally, independently by internal or external reviewers;
- **Learning Loop:** regular reviews of previous evaluations – focusing on content, quality,



and processes – to inform improvements in internal procedures and facilitate capacity building within evaluation teams and among stakeholders.

It is essential that evaluation quality assurance systems incorporate a **dedicated protocol for data quality** addressing accuracy, integrity, completeness, timeliness, and reliability. Continuous real-time data accuracy checks

are recommended to enable immediate corrections as needed, alongside a thorough documentation of all processes related to data collection and handling.

**Transparency** is also critical; all assumptions, methodological steps, and limitations associated with data and analysis must be fully documented and communicated.

---

#### Further reading

Hassnain, H., Kelly, L. and Somma, S. (eds.) (2021), *Evaluation in Contexts of Fragility, Conflict and Violence: Guidance from Global Evaluation Practitioners*, Exeter, UK: IDEAS. Available at: <https://ideas-global.org/wp-content/uploads/2021/06/EvalFCV-Guide-web-A4-HR.pdf>.

Bakrania, S., Balvin, N., Daidone, S. and De Hoop, J. (2021), *Impact evaluation in settings of fragility and humanitarian emergency*, Innocenti Discussion Paper 2021-02, UNICEF Office of Research – Innocenti, Florence. Available at: <https://doi.org/10.18356/25211110-2021-02>.

BetterEvaluation (2024), *Monitoring and evaluation in fragile, conflict-affected and violent (FCV) settings*. [online] Last updated 18 September. Available at: <https://www.betterevaluation.org/methods-approaches/themes/monitoring-evaluation-fragile-conflict-affected-violent-settings>.

## 7

## USAGE OF EVALUATION

Evaluation is a practice- or use-oriented form of social research. Therefore, reflection on **who can use the evaluation, how it can be used, and for what purposes** should accompany the evaluation team from the very beginning of the process. There are many levels and forms of evaluation use, and we should not focus only on one of them as being more desirable than the others.

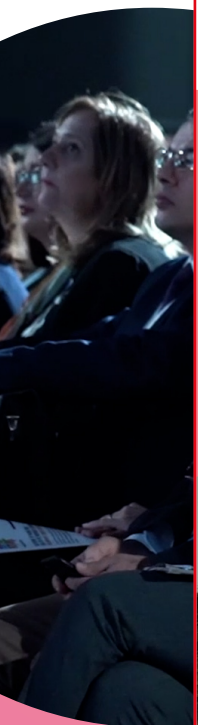
- **Instrumental use** of evaluation occurs when findings and recommendations are directly implemented, for example, by introducing changes in a programme so as to improve its relevance or effectiveness. This type of use is often considered the most critical, as the link between evaluation and its impact is much easier to observe;
- **Conceptual use** relates to the accumulation of knowledge and improved understanding of how programmes or interventions work and what key factors affect their results. For example, programme managers

may better grasp the complexity of an issue, or stakeholders may gain a clearer understanding of what works. This type of use is much more difficult to trace. However, it should not be underestimated as it is crucial for building a better understanding of intervention mechanisms among key stakeholders and, in the long term, may contribute to improved policies;

- **Process use** refers to benefits derived from the evaluation process itself, regardless of its findings. For example, monitoring data may be improved during the evaluation, and staff may learn how to use it more effectively. Different stakeholder groups may also have opportunities to cooperate and better understand their respective perspectives. Although relatively difficult to trace, process use can have a significant positive impact on policy implementation;
- **Symbolic (or legitimising) use** occurs when the fact of an evaluation or its findings are used to justify political or managerial







decisions. For example, a decision to change a programme may already have been made by politicians, but the evaluation is cited as justification. This is often seen as a misuse of evaluation;

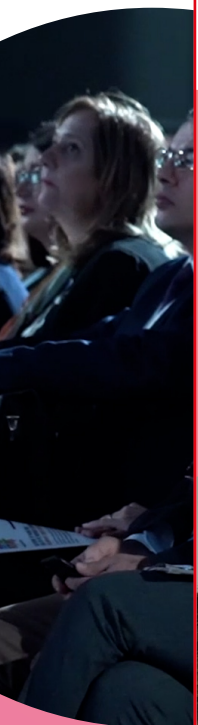
- **Advocacy use** occurs when an organisation employs evaluation findings as evidence in advocacy campaigns or actions aimed at influencing public opinion or policymakers;
- **Accountability use** refers to employing evaluation results to demonstrate compliance, performance, or stewardship to funders, regulators, or the public, for example, presenting evaluation findings to donors or monitoring committees.

All these forms of evaluation use, perhaps with the exception of symbolic use, can be considered beneficial, though in different ways. Each type of use can be critical depending on the vision of the role of evaluation. When evaluation is viewed as part of the policy cycle or managerial systems, **instrumental use** is particularly important. When evaluation is seen from the perspective of knowledge management or learning organisations, **conceptual use** may be more relevant.

The use of evaluation is central to many discussions about evaluation practices and the

institutional contexts in which they take place. Institutional context significantly influences these processes, depending on factors such as the organisation's previous experience with evaluation and the reasons for conducting it – whether due to external requirements from donor organisations, the influence of other organisations considered successful, or an internal belief in the merits of evaluation.

**Involving Evaluation Users.** If an evaluation is to make a real difference, that is to say, that it will be put to some use, any potential beneficiaries must be considered right from the start. Michael Patton (2012), who developed the idea of utilisation-focused evaluation, points out how important it is to identify both potential beneficiaries and the ways they might implement the findings as early as possible. It also helps to bring these users into the process from the beginning, whether that consists of shaping the evaluation questions, setting the criteria and scope, or making key decisions about how the evaluation will be carried out. In addition, potential beneficiaries should be involved in the formulation of recommendations, as they better understand the relevant context and feasibility of any recommendations.



### **Selecting the appropriate timing for evaluation is essential.**

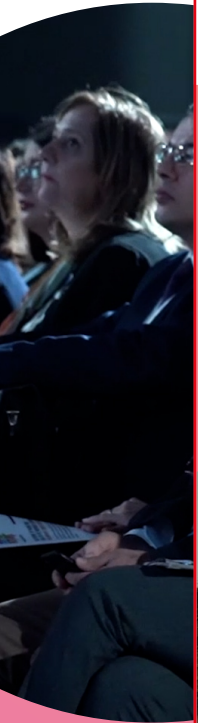
In order to maximise the utility of evaluation outcomes, results should be available when key decisions are being made, such as during the programming or financial planning phase. While meaningful evaluation typically requires some time post-intervention, excessive delays may render reports less effective or obsolete. Since no single approach suits all scenarios, it is important to carefully weigh the advantages and disadvantages associated with each option on a case-by-case basis.

### **Be careful about the main findings and secondary findings of the evaluation.**

Evaluation findings should be properly communicated. The main channel of communication usually consists of an evaluation report, in which evaluation findings are communicated to the main stakeholders. It is likely that the evaluation report will have a limited audience, so it would be beneficial to use alternative methods to disseminate the findings and recommendations to a wider audience.

The [EU guide](#) to better regulation includes a full chapter with detailed recommendations on the evaluation report. The main characteristics are:

- **storytelling** – the report should present a sound, evidence-based narrative on the impact of the intervention, focused on the responses to the critical questions, such as: what was the expected outcome of the intervention? how has the situation evolved over the evaluation period? to what extent was the intervention successful and why? how did the EU's intervention make a difference and to whom? is the intervention still relevant? what are the conclusions and lessons learned?
- **methodological information** – the report should explain what was analysed and how, including information on data and analysis limitations, so as to ensure reliability of both findings and recommendations;
- **a clear chain of logic between the evidence, analysis and findings**, the answers to the questions and any conclusions drawn;



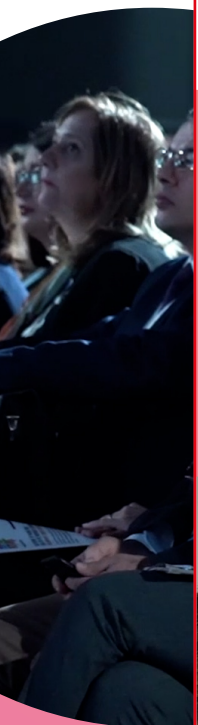
■ **ensuring high quality of recommendations is also critical;** they should not be too generic or cliché. Instead, these should align with the strategic objectives of organisation and intervention. Furthermore, recommendations should be feasible and, as far as possible, concrete and specific. It is a good idea to formulate them in the format ‘who, when and what’, respectively specifying who is responsible, what for and when. The formulation of such a recommendation may require a separate analysis of potential options for solving a given problem, identification of the best, most feasible ways for the solution’s implementation and agreement on the implementation plan. This process usually requires intensive collaboration with key stakeholders.

Besides the report, the publication of other materials communicating evaluation findings should be encouraged, such as an executive summary, presentations, briefings, etc. It is also important to plan communication actions well in advance and to allocate sufficient resources. When, for whom and in which format the dissemination will be conducted

should be clearly defined, e.g. by means of workshops, policy briefs, etc.

**Building a system for tracking recommendations can be useful.** Ensuring that findings and recommendations from evaluations are followed up may be supported by implementing appropriate managerial practices. An example of an advanced management recommendation system is the World Bank’s **Management Action Record**, which annually tracks actions taken in implementing evaluation recommendations. Recommendations are reviewed by the management of the Bank in collaboration with the Independent Evaluation Group, and progress towards achieving intended evaluation outcomes is assessed. The system supports learning, accountability and improvement.

Developing a mature system is not always feasible. However, it is important to integrate feedback from the implementation of evaluation recommendations into management practices and ensure that progress is consistently reviewed by organisational leadership. Without a regular process of review, the influence of evaluation findings and recommendations may be reduced.



### **Integrate evaluation of learning practice within the organisation.**

Evaluation can be used not only to inform policy adjustments but also as a mechanism for supporting organisational development, including for reflection and learning. Incorporating learning outcomes into evaluation questions from the outset can facilitate this approach. Evaluation activities may include opportunities for learning, such as workshops that help analyse results and encourage reflection. This process can contribute to double-loop learning, which goes beyond single-loop learning – where actions are taken based on identified mistakes or areas for improvement – by adding an additional step. In double-loop learning, participants question existing goals, plans, rules, norms, patterns and beliefs (Hummelbrunner, 2013), thereby making discussions of evaluation findings more comprehensive, and potentially leading to changes in foundational organisational mechanisms.

### **Some organisations establish evaluation brokers**

– either internal or external intermediaries – who facilitate the translation of evaluation findings for users according to their specific needs. These brokers operate between those generating evaluations and those utilising

them, particularly when users may lack the time, resources or expertise to analyse and apply findings independently. Brokers play a crucial role in convening relevant stakeholders at appropriate junctures to discuss evaluation outcomes or contribute to policy development based on collected evidence. Additionally, they assist in assessing the quality and reliability of evaluation results. Evaluation brokers can also enhance capacity on both the producer and user sides, fostering improved alignment.

Brokers may serve as internal units, such as an ‘Evidence & Learning Hub’, or as external entities. A notable example of the latter is the [What Works Network](#), a consortium of independent organisations specialising in identifying, analysing and assessing impact evaluations in targeted areas, and formulating policy guidelines grounded in evaluative evidence.



## 8

## STRENGTHENING INSTITUTIONS TO PROPERLY DESIGN AND USE EVALUATION

A single evaluation rarely influences institutional actions unless it is integrated into the broader management and learning systems of the organisation. It is based on the assumption that effective public policy requires solid data and evidence to make evidence-based decisions. Such information can only be provided by reliable monitoring and evaluation systems, highly integrated with other performance management tools, such as budgeting, planning, reporting, assessment and policy revisions.

**Monitoring and evaluation systems may be defined as the set of institutions, functions, processes** and resources that enable governments and stakeholders to systematically collect, analyse and use evidence on the implementation, performance and results of policies, programmes or projects. In fragile contexts, the system must also be feasible, flexible and support institutional trust. It should recognise that access to data can

be limited, the capacity of institutions and individuals can be weak, and that conditions may change rapidly.

The significance of a monitoring and evaluation system lies in its capacity to consistently and systematically collect, analyse and utilise data and evidence. In this context, evaluation becomes an integrated, ongoing process rather than an ad hoc or isolated activity conducted by external stakeholders, such as donors. It represents a structured approach to providing reliable information for decision-making, necessitating established frameworks, qualified personnel, defined processes and appropriate resources.

The development of a Monitoring and Evaluation (M&E) system should commence with a thorough assessment. One example of an assessment framework for establishing an M&E system is the [Diagnostic Tool for a Monitoring and Evaluation Systems Analysis](#)





(MESA), developed by the [Global Evaluation Initiative](#). This tool facilitates the review and evaluation of the capacity of the monitoring and evaluation system, assists in identifying gaps and potential areas for enhancement, and serves as an effective starting point for further development and strengthening of the system.

The conceptual framework for assessing the capacity of a monitoring and evaluation system comprises three main elements:

- **enabling environment:** this encompasses the social and institutional systems of the country, including its laws, regulations, policies, power structures and prevailing social norms. Crucial factors within this component include the political context, cultural environment, incentives provided and the general understanding of M&E. These aspects can promote the stable evolution of evaluation practices, especially when there is encouragement to conduct and utilise evaluations in policy processes. Conversely, a discouraging environment, such as limited tradition or capacity for evidence-based policymaking, may hinder progress;
- **organisational capacity development:**

this refers to the ability to manage and coordinate a comprehensive M&E system. The presence of robust national planning or strategic bodies, such as the ministry of finance or ministry of development, may enhance government capacity. Alternatively, the system can operate at the sectoral or sub-national level, for instance, within individual ministries. Assessments focus on the roles and responsibilities of key stakeholders concerning monitoring and evaluation, as well as coordination and cooperation mechanisms;

- **individual capacity development:** given that the system is operated by individuals, its effectiveness depends not only on well-designed structures and procedures but also on committed and well-trained personnel. These individuals and champions are essential for establishing and maintaining the appropriate components of the system.

The assessment of the evaluation and monitoring system is only the starting point. Building or strengthening the system requires much more effort, planning and actions. Especially in fragile contexts it requires a phased, adaptive and inclusive approach. Traditional blueprints often fail in environments



characterised by political instability, weak institutions, and limited data infrastructure. The following recommendations highlight practical steps for strengthening M&E systems in such settings.

**Begin with a practical, sufficiently robust system** – in fragile contexts, it is advisable to avoid excessive investment in complex and resource-intensive monitoring and evaluation (M&E) systems, as this may not yield the desired outcomes. It is recommended to implement smaller and simpler solutions, such as basic monitoring tools that provide key information and straightforward evaluation methods. These initial systems can be expanded and enhanced over time as organisational and individual capacity increases.

**Prioritise organisational capacity** – the system requires appropriate structures, procedures and resources. It is strongly recommended to establish a dedicated unit for monitoring and evaluation, staffed with committed and well-prepared professionals. Investing in the capacities and skills of those who manage and operate M&E systems is essential. Skills development should focus on practical training, mentoring and peer-to-

peer exchange with stakeholders. Conducting monitoring and evaluation in practice remains the most effective way to learn, provided it is complemented by sufficient opportunities for reflection and knowledge sharing. Strengthening cooperation with stakeholders is also crucial, including other units within public administration, non-governmental organisations, researchers and consultants, the donor community and statistical offices.

**Embed M&E within national systems and strategies** – one of the main weaknesses of the M&E system can be its fragmentation, i.e. the co-existence of independent systems of monitoring and evaluation in a given field. This type of situation often occurs when stakeholders do not cooperate sufficiently, and donors establish their own monitoring and evaluation requirements. To avoid such situations, M&E activities should be aligned with national strategic and planning mechanisms and public statistic systems. In addition, coordination and harmonisation of donor activity is required, as it can reduce duplication and strengthen national ownership.



**Strengthen the capacity of the M&E system to produce robust data**

– one of the key elements of the monitoring and evaluation system is ensuring access to robust, high-quality monitoring data for programmes and policies. Therefore, strengthening evaluation systems requires investment in improving the availability and reliability of monitoring, usually administrative data. This type of data can serve as a basis for more advanced impact evaluation. This type of investment is critical in fragile contexts, where limited access to data is highlighted as one of the main obstacles to evaluation.

**Use mixed methods and innovative data sources**

– national administrative data systems may have some gaps. For this reason, it can be beneficial to use different, qualitative and quantitative methods of data collection, such as household surveys, perception studies, participatory monitoring and case studies. It can complement official statistics and provide for a richer understanding of policy results. New technologies such as mobile phone-based surveys, geospatial data, or satellite imagery can also help overcome barriers to traditional data collection.

In order to make evaluation an effective policy improvement tool, the focus must not only be on promoting evaluation itself but also on strengthening the institutional context in which it occurs. Evaluation tends to be more impactful when supported by a systemic approach with suitable structures and capacities.



## 9

## REFERENCES

- ADLS (2016) *Administrative data introduction*. Available at: <http://www.adls.ac.uk/adls-resources/guidance/introduction/>
- Brenner, M. E. (2009), 'Interviewing in Educational Research', in Green, J.L., Camilli, G., Elmore, P.B. (eds.) *Handbook of Complementary Methods in Education Research*, New York: Routledge.
- Card D, Ibarra P, and Villa J M, (2011), *Building in an Evaluation Component for Active Labour Market Programmes: A Practitioners Guide*, IZA DP No 6085, IZA, Bonn. Available at: <https://ftp.iza.org/dp6085.pdf>
- Davies, H. T. O., Nutley, S. M., Smith, P. C. (eds.) (2000), *What Works? Evidence-based Policy and Practice in the Public Services*, Bristol: Policy Press.
- ETF (2025), *Evaluating Active Labour Market Policies in Fragile Contexts*, ETF, Turin. Available at: <https://www.etf.europa.eu/en/publications-and-resources/publications/evaluating-active-labour-market-policies-fragile-contexts>
- Hummelbrunner, R. and Reynolds, M. (2013), 'Systems thinking, learning and values in evaluation', *Evaluation Connections: The European Evaluation Society Newsletter*, June, pp. 9–10.
- Hutchinson, K. (ed.) (2019), *Evaluation Failures: 22 Tales of Mistakes Made and Lessons Learned*, Thousand Oaks: SAGE Publications.
- ILO (2018), *Guide on Measuring Decent Jobs for Youth - Monitoring, evaluation and learning in labour market programmes. Note 5: Identifying an appropriate impact evaluation method*, International Labour Office, Geneva.
- Kitchin, R. (2014), *The Data Revolution*, Washington, DC: Sage Publications.
- OECD (2022), *States of Fragility 2022*, OECD Publishing, Paris. Available at: <https://doi.org/10.1787/c7fedf5e-en>
- Patton, M.Q. (ed.) (2012), 'Special issue: Stories of partial success', *Canadian Journal of Program Evaluation*, 27(3).
- Smith, N., Noble, M., Antilla, C., Gill, B., Zaidi, A., Wright, G., Dibben, C. and Barnes, M. (2004), *The value of linked administration records for longitudinal analysis*, Social Disadvantage Research Centre, University of Oxford.
- Veltri, G.A. (2011), 'Delivering Effective Qualitative Research for Policy Making', presented at Qualitative Research for Policy Making conference, Belfast, 26-27 May.
- Veselý, A., Nekola, M. (eds.) (2007), *Analýza a Tvorba Veřejných Politik; přístupy, metody a praxe*, Praha: SLON.





European Training Foundation

-  [www.etf.europa.eu](http://www.etf.europa.eu)
-  <https://bsky.app/profile/etf.europa.eu>
-  [www.youtube.com/user/etfeuropa](http://www.youtube.com/user/etfeuropa)
-  [www.facebook.com/etfeuropa](http://www.facebook.com/etfeuropa)
-  [www.instagram.com/etfeuropa](http://www.instagram.com/etfeuropa)
-  [openspace.etf.europa.eu](http://openspace.etf.europa.eu)
-  <https://www.linkedin.com/company/etfeuropa/>

ISBN 978-92-9157-761-3