European Commission

DESIGN AND COMMISSIONING OF COUNTERFACTUAL IMPACT EVALUATIONS

A PRACTICAL GUIDE FOR ESF MANAGING AUTHORITIES

Employment, Social Affairs and Inclusion

EUROPEAN COMMISSION

Directorate-General for Employment, Social Affairs & Inclusion
Directorate G — Funds, Programming and Implementation
Unit G5 — Better Regulation

Contact: Linda Adamaite

E-mail: EMPL-G5-UNIT@ec.europa.eu or Linda.Adamaite@ec.europa.eu

European Commission
B-1049 Brussels

EUROPEAN COMMISSION

DESIGN AND COMMISSIONING OF COUNTERFACTUAL IMPACT EVALUATIONS

A PRACTICAL GUIDE FOR ESF MANAGING AUTHORITIES

Directorate-General for Employment, Social Affairs & Inclusion

2021

Manuscript completed in September 2021

The 2021 edition of “Design and commissioning of counterfactual impact evaluations - A practical guide for ESF managing authorities” has been authored by Jochen Kluve (Humboldt-Universität zu Berlin) and Andrea Naldini and Marco Pompili (Ismeri Europa). This is an adaptation of the 2013 edition of the guidance (ISBN 978-92-79-28238-6; DOI 10.2767/94454) prepared by Stephen Morris (Policy Evaluation and Research Unit, Manchester Metropolitan University), Herta Tödtling-Schönhofer (Metis GmbH, Vienna) and Michael Wiseman (George Washington Institute of Public Policy).

Luxembourg: Publications Office of the European Union, 2021

PDF ISBN 978-92-76-40725-6 doi:10. 10.2767/02762 KE-02-21-992-EN-N

Contents

LIST OF BOXES

LIST OF FIGURES

LIST OF TABLES

Introduction: background and purpose of the Guide

A guide for managing authorities updating the 2014-2020 experience

The 2021-2027 programming period begins with the dramatic experience of the COVID-19 pandemic and the subsequent economic crisis. The European Union (EU) increased its financial and policy efforts to aid the recovery of national economies and increase employment. In this framework, the European Social Fund plus (ESF+) plays a key role in delivering widespread assistance to the unemployed, with a focus on young people and women, as well as supporting interventions against child poverty, promoting better education and social inclusion of weaker social groups across the EU. The diverse goals pursued by the ESF+ and its need to achieve rapid results in the areas of employment and social inclusion, require an effective allocation of resources. Evidence-based approaches to policy-making are ever more important, and evaluation is a fundamental instrument to direct public policy.

In the 2014-2020 period the European Commission (EC) supported the evaluation capacity of Member States (MSs) and managing authorities (MAs), and promoted the use of the Counterfactual Impact Evaluation (CIE). The positive results found in the CIE are tangible evidence of ESF effects beyond what would otherwise have been achieved. A significant number of CIEs were first scheduled in the Evaluation Plans (EPs) and then carried out during the implementation of ESF programmes. In many cases CIEs encountered difficulties during their preparation and implementation, or remained isolated experiences and were not included in a systematic evaluation framework.

This Guide is intended for managing authorities (MA) and other bodies responsible for the implementation of ESF+-funded interventions and programmes and aims to aid the planning, design and commissioning of CIEs. It takes into consideration prior experience and provides practical advice on some of the key questions to be considered in developing a CIE. The Guide updates the previous 2014-2020 Guidance, placing more emphasis on the issues encountered in the practical implementation of a CIE. Nevertheless, methodological aspects are discussed and, where possible, simplified and integrated with updated examples selected from 2014-2020 ESF evaluations.

A CIE addresses 'what works?'

CIEs address the crucial questions that enable evidence-based policy decisions: what are the causal effects of interventions and 'what works?'' They seek evidence of whether ESF-financed interventions are actually responsible for the changes in participants' circumstances and the consequent achievements of the interventions. When executed well, CIEs provide evidence of the 'net effect'[1], or impact, of an intervention, enabling policymakers to rule out alternative explanations for changes in participants' circumstances or accomplishments that may have been observed. When CIEs provide estimates of the presence and magnitude of the net effect,

[1] The net effect, or impact, is the residue between the total (or gross) effect and what would have been achieved in the absence of the intervention. The net effect may also be negative, when the intervention is less effective than the market dynamics. See p.9 for further details.

these necessarily contain a measure of uncertainty, depending on the methodological accuracy and available information. The type of evidence provided by CIEs enables policymakers to assess the effectiveness of interventions, make comparisons between interventions and assess their relative performances. Furthermore, it provides important inputs into cost-benefit or cost-effectiveness analyses.

This Guide is published at a time of unprecedented challenges for ESF+. Given the huge increase in EU resources for investments and employment provided by the Next generation EU package, it is critical that policymakers measure and understand the effects of the interventions they are responsible for. Public funds must be allocated to more productive and effective interventions to speed up recovery and reduce social imbalances. It is, therefore, incumbent upon those responsible for disbursing ESF+ resources to justify their choices by demonstrating that their interventions are effective and provide value for citizens. The best way to achieve this is by conducting a greater number of high-quality CIEs.

The ESF+ is the main European instrument for supporting employment and social inclusion. In the programming period 2014-2020, the ESF spent nearly € 125 billion on active labour market. education and social inclusion policies implemented through operational programmes (OP) in the 28 Member States. As stipulated in General Provisions Regulation 2013/1303, evaluations 'shall be carried out to improve the quality of the design and implementation of the programmes, as well as their effectiveness, efficiency and impact'.

Results orientation and high-quality evaluation

In the programming period 2021 – 2027, performance and results will continue to be under examination[2]. This will require reinforcement of current monitoring and evaluation systems and capacities, including data collection arrangements. Evaluation plans will remain obligatory, and further emphasis will be placed on impact evaluation. As a variety of methods are available to capture the impacts of ESF+ supported operations, the managing authorities must decide which method, or which combination of methods, is most suitable to satisfy the regulatory requirements. A rigorous quantification of impacts of interventions also involves counterfactuals.

The focus on strong performance and result orientation is an important feature of the new regulations. High-quality evaluation strategies and techniques are essential to acquire essential knowledge showing all MSs which interventions 'work' and which do not. Strengthening the quality of evaluations and developing reliable evidence of value added is essential.

In principle, the starting point for gathering evidence on the effectiveness of policy interventions is straightforward. The requirements include:

- Identification of the problem to be addressed
- Identification of the instruments to be employed to address the problem

A formula that connects the instruments and the results.

[2] REGULATION (EU) 2021/1060 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 24 June 2021 laying down common provisions on the European Regional Development Fund, the European Social Fund Plus, the Cohesion Fund, the Just Transition Fund and the European Maritime, Fisheries and Aquaculture Fund and financial rules for those and for the Asylum, Migration and Integration Fund, the Internal Security Fund and the Instrument for Financial Support for Border Management and Visa Policy

The need for results...

In order to evaluate a funding scheme or instrument it is fundamental to have clear and measurable indicators of both the inputs applied and problem-related outputs and results. It is common to set targets for both outputs and results, and to compare actual achievements to targets. Monitoring is employed to track inputs, outputs and results over time, and provide management feedback. The underlying intervention theory often points to intermediate results that may also become the focus of monitoring. But getting from here to identifying whether a particular intervention works is not simple.

...and evidence of net effects

There is a need to supplement existing evaluation practice with approaches that generate much stronger evidence of the net effects or impacts of interventions. Measuring what is achieved is a matter of accountability for funds used. CIE addresses the fundamental question of whether an intervention is effective. While CIE attempts to establish a causal link between interventions and results, further theory-based and process evaluation methods may be required to identify the underlying causal mechanisms and to help ensure that impacts attained in one location provide an evidence base for policy replication elsewhere.

MS experience with CIE

In the 2014-2020 programming period all Member States and managing authorities adopted an evaluation plan describing evaluation objectives, activities, organisational elements and products. The evaluation plans envisage several types, such as overall on-going evaluations of the programmes, and also thematic evaluations aimed at answering specific evaluation questions or focusing on specific axes or investment priorities of an OP. An analysis of the evaluation plans relating to 177 ESF programmes, as at the end of 2018[3], showed that around 132 counterfactual impact evaluations are expected in this programming period, a figure which indicates progress and greater focus on the counterfactual methods in the 2014-2020 programming period compared to 2007-2013[4].

The use of CIE is increasing

According to the Evaluation Helpdesk projects repository[5] 1,795 evaluations were completed up to June 2021, out of which 1,001 related to ESF/YEI or ESF/YEI and ERDF programmes (675 and 326 respectively). Of 1,001 evaluations covering ESF, 323 were impact evaluations and 106 of them applied a counterfactual approach: 82 out of 234 impact evaluations related to ESF/YEI programmes and 24 out of 89 impact evaluations were of ESF/YEI and ERDF programmes[6].

In the first years of the current programming period counterfactual evaluations of ESF/YEI and/or ERDF programmes focused on the previous programming period, while since 2018 almost all CIEs assess the effects of the 2014-2020 interventions[7]. 20 out 27 Member States produced counterfactual evaluations of programming including ESF, reflecting also in this case a more widespread adoption of this approach compared to the past.

[3] See Ismeri Europa – Ecorys – Institute for Employment Studies, 2019.
[4] See Bratu C. et al., 2014
[5] The Helpdesk project, funded by DG REGIO and DG EMPL, collects information on the evaluations published online since the 1st of January 2015 on the websites of MAs. The identification of the evaluations is carried out by a network of national experts and they are summarized and assessed in terms of quality and reliability of findings.
https://ec.europa.eu/regional_policy/en/policy/evaluations/member-states/
[6] 90 out of 106 were examined and summarised by the Helpdesk project.
[7] 29 of 33 CIEs carried out up to 2018 relate to the previous programming period.

Although methodological approaches to CIEs vary across Member States, the counterfactual approach was most commonly implemented to assess the impact of active labour market policies (training, incentives, job search assistance, work experiences) but to a much lesser extent in other fields such as interventions to support educational institutions and students, and policies relating to social issues, such as interventions to fight poverty or social exclusion.

In methodological terms, the propensity score matching technique is the most frequently used, while other methods are less prominent. CIEs often focus on short- or medium-term effects (6 or 12 months) while long-term effects are less frequently assessed.

But limitations are still frequent

The analysis of CIEs carried out in the Evaluation Helpdesk project showed limitations such as: people in control groups not similar enough to those in the support groups, low reliability of data, for example when drawn from ex-post surveys, interviews relying on respondents' ability to recall events accurately, or small sample size producing statistically insignificant results, etc. In some cases, the evaluation reports do not contain sufficient information on the methodological steps, choices and limitations, which are necessary elements to inform others, apart from those who commissioned them, on the effects of the measures examined, adding to their knowledge of the effects.

A Guide for practitioners

To sum up, despite progress in recent years, the execution of CIEs demands technical expertise and political will. This Guide makes the case for CIEs, and sets out some of the issues that MAs need to address to make their implementation successful. Beyond the practical aspects of CIEs, there is also focus on the wider issues that may need to be addressed to enable impact evaluations of higher quality. The Guide includes a number of basic recommendations, hopefully helpful for the MAs, but one of the main messages is that carrying out CIEs requires careful planning of data preparation (especially administrative data), clear aims and objectives, etc. in order to avoid potential problems in the implementation phase or low-quality evaluations.

The Guide also includes examples of evaluations and practices to aid the reader. While the evaluations collected by the Evaluation Helpdesk project provide the main source of information to identify the examples and practices presented in this Guide, other sources were also considered, especially the evaluations carried out by the Joint Research Centre (Centre for Research on Impact Evaluation (CRIE)) and to a more limited extent, academic publications. Where possible, examples and experiences included in the Guide relate to European Social Fund interventions.

The Guide is structured in four chapters.

A structure reflecting the steps in an evaluation

Chapter 1 discusses the nature of CIEs and why they are important. It provides an introductory overview of CIE approaches, emphasising the distinction between experimental and quasi-experimental approaches. It also introduces the technical and practical challenges that need to be tackled when implementing a CIE. Consideration is given, in general terms, to the types of policy questions that might be addressed by CIEs and the relationship between CIE methods and other approaches to evaluation (for example: theory-based approaches, process evaluation and efficiency analysis).

Chapter 2 looks at a series of questions that MAs should consider when designing and implementing CIEs. This Guide sets out some of the key challenges that commonly confront those developing CIEs and makes some recommendations as to how these might be addressed. The questions are a guide for those aiming to commission CIEs of ESF-financed interventions.

CIEs provide robust evidence of the effectiveness of funds. They only do so, however, if they are well planned and executed appropriately. In order to do this, MAs have to address certain key issues in commissioning an evaluation. The precise manner in which MAs consider these issues and the order in which they do so, will be dictated by practicalities and institutional arrangements on the ground in the Member State. This Guide highlights some of these important issues and draws them to the attention of MAs. Main issues are listed according to six steps of the evaluation process: 1) Selecting the operation(s) for assessment; 2) Identifying the evaluation questions and the outcome variables; 3) Analysing data to identify the control group and to measure the outcome variables; 4) Selecting the CIE method; 5) Defining the timetable and the budget; 6) Implementing the evaluation. Particular attention is dedicated to data availability and protection, an issue which can seriously compromise the possibility of carrying out a CIE.

Chapter 3 explores step 4 above (Selecting the CIE method) and focuses on the characteristics of the main methods used in counterfactual analyses. In particular, five methods are presented: the experimental method or randomisation, propensity score matching, difference-in differences, regression discontinuity design and instrumental variables. This section of the Guide does not intend to train the readers to the use of the CIE methods, but aims to make readers aware of the strengths and weaknesses of each method and where one method is more suitable than another.

Chapter 4 addresses wider issues of perspective development. These include the need to develop capacity to conduct CIEs successfully, both within MAs (policy makers and officials) and among MSs' research, academic communities and consultancy companies. This section also addresses the need to confront legal barriers around data access and update the CIEs of ESF+ programmes to more advanced designs to extend coverage of social inclusion and education policy, the estimation of effects for "soft outcomes" and the implementation of meta-evaluation approaches.

In sum, this Guide: 1) makes the case for CIEs, 2) identifies the important steps towards successful execution of CIEs and 3) sees CIEs as an essential part of the ESF+ landscape. The ultimate objective is to enhance the contribution of ESF+ to the well-being of Europe's citizens.

The authors wish to acknowledge the valuable support and assistance provided by members of the Unit G5 of DG Employment, Better Regulation in producing this report, in particular, Linda Adamaite Jeannette Monier and Maria José Cueto Faus. The authors also wish to thank Andrea Pisano and Ernesto Belisario for their contribution to the section on data protection and to Paweł Hess for his support and contribution.

Chapter 1. Concept and approaches

This chapter addresses fundamental questions about the nature and the purpose of the counterfactual approach in assessing causal effects of interventions. Specifically, it sets out an understanding of the essence of the counterfactual impact evaluation, particularly as it relates to the types of operations co-financed through ESF+. It also examines the relationship between counterfactual approaches and other evaluation methodologies and discusses why CIEs are important - particularly at the present time. The policy issues that CIEs can address are examined, and a brief, overview of some of the main counterfactual methods relevant to evaluating ESF+ co-financed interventions are introduced before being presented in detail in Chapter 3.

1.1. Essence of the counterfactual

CIEs seek to identify the net effects or impacts of interventions. Their distinctive feature is that they aim to support claims that a given intervention causes a specific result; that is, the specific result can be attributed solely to the intervention. CIEs achieve this by isolating the intervention and ruling out alternative explanations for the observed result.

The counterfactual

Underlying their capacity to rule out alternative explanations is the idea of the 'counterfactual', that is, the answer to the question "What would have happened – in terms of the outcomes of interest – if the intervention had not been implemented?". To understand clearly the concept of the counterfactual, and put very simply to clarify the issue, it is helpful to consider the example of an unemployed individual participating in a training programme, the aim of which is to encourage employment. In order to determine the effect of training on the individual, the counterfactual approach conceives of two potential results[8]. The first is the trainee's employment status subsequent to having taken part in training. This is the observed result for the trainee. The second potential result is this trainee's employment status had he or she not taken part in the training programme, all else being equal. In these circumstances this second result is referred to as the counterfactual result. The impact of training for the individual trainee is identified by the difference between the observed and counterfactual results. This is the causal net effect or impact of the training for the individual. The only difference between the circumstances or conditions which gave rise to the observed and counterfactual results is the individual's participation in the training. Therefore, any difference between the two results must be the impact of training on the individual's employment status.

Defining treatment groups and ....

In reality we do not and cannot observe counterfactual results for individuals exposed to an intervention. The chief aim of CIE, however, is to provide convincing estimates of counterfactual results for groups of individuals or enterprises affected by ESF co-financed interventions. Thus, impacts are expressed, for example, in the form of differences in means or proportions between average observed and 'estimated' counterfactual values. In most applications, CIEs seek to compare the results of an intervention (a measure or an instrument) for those entities (persons, SME, etc.) that benefitted from

[8] A more detailed discussion of the 'potential outcomes' model of causation can be found in Holland P., 1986.

it to a group not subject to the intervention. In the terminology of CIE, the 'treated' or 'treatment' group is distinguished from the 'control' group, which should be as similar as possible in all respects (except for the treatments being received) to the treated group. It is from the control group that estimates of counterfactual results are obtained, with specific attention paid to differences in characteristics - observed and unobserved - between the two groups. It is also possible to compare a number of different treatments by exposing eligible units to a range of treatment variants (e.g., other ESF-funded treatments or interventions funded through other sources), forming a number of treatment groups and comparing results one to another, and/or results for a non-treated control group.

... control groups

Where the control group is exposed to no treatment, the evaluation question addressed is 'What is the impact of receiving the intervention relative to receiving no help or support?' Conversely, where the results of receiving the treatment of interest are compared to the results of receiving some other treatment, the evaluation question addressed is: 'What is the impact of receiving the intervention under consideration relative to being exposed to some well-defined alternative?' A CIE can in many cases be designed to address either of these fundamental questions. The choice of which question to address is determined by policy makers' priorities and practical design constraints.

In cases where a comparison is made between two different treatments, there should be a clearly defined contrast between them, which is meaningful from the perspective of policy making.

1.2. Why are counterfactual evaluations important?

CIEs provide important information about the net effects, or impacts of interventions. They provide estimates of the magnitude of impacts, their sign (whether positive or negative) and statistical measures of uncertainty. They help to validate or reject the presumed causal connection between the intervention and results, that underlies the design of the intervention. These measured effects can be used in pursuing a number of aims: to demonstrate transparency and accountability in spending taxpayers' resources; to take policy decisions based on clear evidence; to learn from results across regions of the same country or across MSs.

Transparency and accountability

Those responsible for interventions and concerned with ensuring that their programmes continue to attract funding will have a keen interest in promoting CIEs in order to show that their programmes provide value for money and yield measurable benefits to participants, as well as to society as a whole. Evidence from CIEs will be of particular interest to those responsible for resource allocation. MAs will be eager to show that their programmes do indeed 'work'. To do this convincingly, they will need to commission high-quality CIEs.

Supporting resource allocation decisions

These features of CIEs provide important information to policy makers whose task it is to allocate resources to different interventions. Decisions regarding the funding of potential interventions take place within a context of resource limits. In this context, decision makers need sound evidence of programme impacts and cost effectiveness so they can use the available resources to best effect. In the assessment of the relative efficiency of interventions, the

net effects estimated by CIEs can be displayed in greater detail by studying an intervention's cost effectiveness or undertaking a full cost-benefit analysis.

Comparing operations of different regions or MSs

Important additional knowledge can also derive from the comparison of the net effects of similar operations implemented in different territories. If CIEs detect significant differences in their effects, it very likely means that the quality of the operations differs significantly or that some operations fit their socio-economic contexts better than others. These comparisons - generally called 'meta-evaluations' or 'meta-analyses' - allow a higher generalization of the CIEs' findings and provide important tests for policy measures (e.g., training, and integrated services, etc.)[9]. To carry out these comparisons according to scientific criteria, an adequate number and a systemic implementation of CIEs are necessary. Consequently, good coordination at national or EU level is a fundamental condition to make these comparisons more frequent and of use in informing policy decisions.

1.3. Why are counterfactual evaluations technically challenging?

Moving beyond simplistic approaches

There are a number of approaches that might be described as 'unreliable' attempts at estimating intervention impacts. These are discussed here in order to illustrate the complexities inherent in CIEs and no reference is being made to actual evaluation practice.

First, a policymaker may wish to evaluate the impact of a training programme for the unemployed by comparing income for trainees subsequent to training, with income for all unemployed persons who did not participate in the programme. The policymaker then attributes to the training programme the observed difference in income between participants and non-participants.

This is not a valid strategy for identifying the impact of training on income, because non-trainees may differ in important ways to trainees, and these differences may influence results - therefore, such an approach does not rule out alternative explanations for any differences in income observed. For example, trainees may have greater inherent ability than non-trainees. In other words, unemployed persons of greater ability volunteer to participate in the training programme. Thus, ability does not only affect the decision to participate but also the results - unemployed persons with higher levels of inherent ability are more likely to command a higher income than those with lower ability. As a result, any observed difference in income between treatment and control groups may be due to both the training programme and/or inherent differences in ability, and one would thus 'over-estimate' the impact of the intervention.

Counteract selection bias

If ability cannot be measured and differences in inherent ability between the two groups cannot be taken into account when estimating impacts by comparing outcomes between the two groups, then the estimated impact of the training programme is said to suffer from **selection bias.** To address this problem, evaluators attempt to collect as much information as possible on important factors that affect the decision to participate and the resulting outcomes. These data are employed to select a valid control group from non-participants– that is, a group that is as similar as possible to the treatment

[9] See, for instance, Card D., Kluve J. and Weber A., 2017.

group in terms of these factors – and conduct statistical analyses taking these factors into account. In doing so, evaluators often invoke the assumption that selection into the programme is determined by these observable factors. This 'identifying assumption' cannot be easily validated in general, and the evaluator needs to argue convincingly that the assumption is plausible in the given context on the basis of knowledge of institutional factors and behavioural theory.

Before and after change

A second 'unreliable' approach might be for the policy maker to observe income for trainees before and after training, and attribute the before/after change to the training intervention. In essence, this approach assumes that in the absence of the intervention average income remain unchanged.

Again, in almost all cases this is not a valid strategy for measuring the impact of training on income, unless the assumption of temporal stability can be plausibly invoked. This is because trainees' income will inevitably change over time in ways that are completely unrelated to training. For example, it is common to observe that the earnings of trainees dip prior to participation, partly due to transitory factors. In many cases rebound would occur regardless of a training intervention.[10]The unreliable approach of gauging the impact of training by the difference between earnings immediately before programme entry and earnings afterwards ignores the fact that, in many cases, earnings would have risen anyway.

To adjust such designs, a measure of the counterfactual - that is, a measure of how income would have changed for trainees in the absence of the training intervention - is required. For example, such a counterfactual result can be obtained from a carefully matched control group not exposed to the training intervention and whose incomes are observed at the same points in time as those of the trainees. The common trends assumption is then often invoked, which posits that the trend in incomes among trainees and the control group would have been the same in the absence of the intervention.

The limits of these 'unreliable' approaches motivate the search for more convincing methods of evaluation. As has been suggested above, more convincing methods are, however, more technically challenging to implement. The next section of this chapter provides a brief outline of some of the specific approaches to CIE that are likely to be most relevant in an ESF context.

1.4. An overview of CIE designs and approaches

When focusing on the effects of an intervention on the participants, counterfactual results are usually estimated using data collected from groups of non-participants who are similar to those participating in the intervention being evaluated. Table 1 at the end of this chapter presents a brief overview of approaches, some of their advantages and limitations, and the essential types of data they require.

The main distinction in CIE is between evaluation designs that are experimental and those that are quasi-experimental. The experimental approach is commonly referred to as the 'randomised control trial', or RCT, and sometimes also as 'social experimentation'.

[10] This pattern is famously called the 'Ashenfelter Dip' after the economist who first commented on it. See Ashenfelter O., 1978.

Randomised design - the golden standard

The experimental approach is considered the golden standard among CIE methods for evaluating the effects of interventions that can be tested and manipulated over relative short time spans, and in most circumstances represents the ideal. A good impact evaluation design should strive to obtain estimates of counterfactual results that are unbiased. In many applications, an experimental approach can be considered as yielding such unbiased estimates. In discussing approaches to CIE, it is often desirable to start by outlining the experimental approach. This is because quasi-experimental methods essentially seek to mimic the experimental ideal.

In discussing CIE designs, the key features of each approach are set out as simply as possible in order to clarify the underlying principles. In reality, applications of these methods can be considerably more complex, and issues such as non-compliance - i.e., when individuals assigned to participating in the intervention did not participate - can add significantly to the challenges faced.

Non-randomised or quasi-experimental designs

There are a wide range of approaches that essentially seek to mimic randomisation. These are referred to as being **quasi-experimental.** It is not possible to review them all within the confines of this Guide, or to provide a complete, detailed technical account of each one. However, in broad terms the quasi-experimental methodologies most likely to be implemented in the context of the ESF are: 1. propensity score matching; 2. difference-in-differences; 3. regression discontinuity; 4. instrumental variables. An overview of major approaches and their relative merits is provided in Table 5 in Chapter 3; their presentation is greatly simplified in order to highlight the key principles of each approach. Further readings on quasi-experimental methodologies are presented in Annex 1.

1.5. How CIE can be embedded in a wider evaluation framework

Causal explanation and description

Counterfactual evaluations address certain types of questions about the causal effects of interventions. These approaches are constrained to the extent to which they might address other questions regarding an intervention. It is helpful to distinguish between evaluation questions concerning **causal explanation** and those regarding **causal description.** CIEs aim to **describe** the consequences of an intervention. Such methods are less suited to **explaining** the mechanisms and contexts through which causal relationships arise. This distinction is an important one, as it helps to clarify the distinctive role of CIE[11].

What CIEs can tell policy makers and what they cannot

A well-designed CIE will tell the policy maker whether an intervention has led to the change in results it was designed to influence. It will provide evidence of the size of any impact, or effect, tell the policy maker whether the impact was positive or negative, but also provide a measure of uncertainty. What counterfactual impact evaluations do less well, is to provide an account of why and how the impacts that are measured through the CIE came about. Conversely, it is often difficult to determine, on the basis of a CIE, why an intervention had no impact, if that proves to be the case.

Within most policymaking bodies, the stakeholders asking causal descriptive and causal explanatory questions tend to have different interests and

[11] See Shadish W.R., Cook T.D. and Campbell D.T., 2002, and Stern E. et al., 2012.

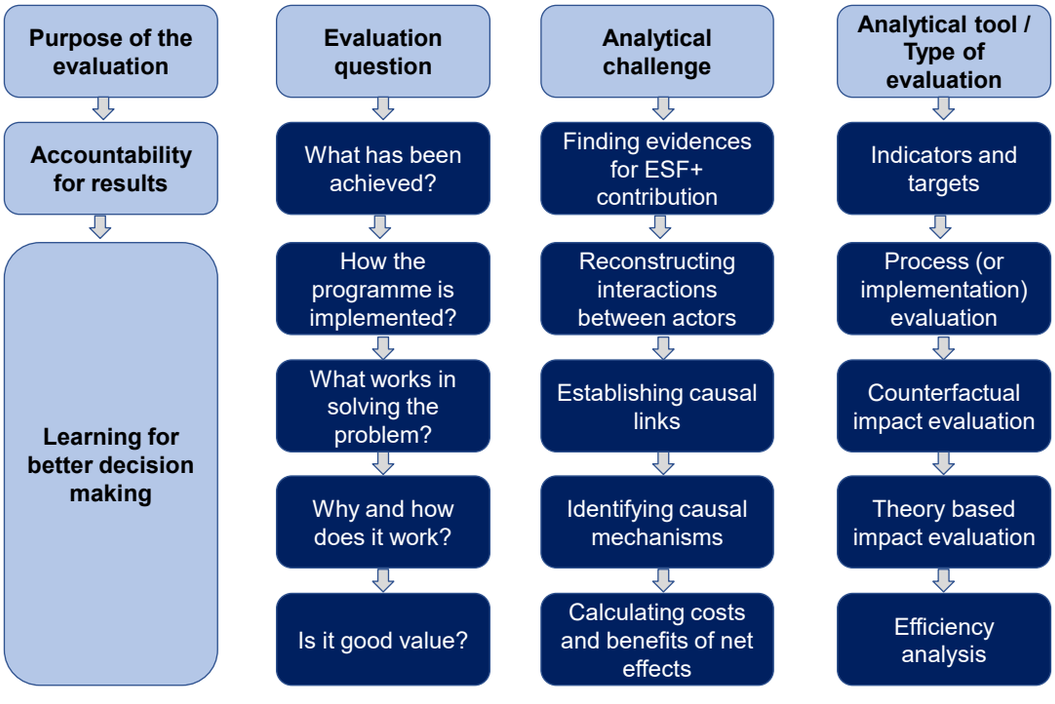
perspectives. Programme managers and practitioners tend to focus on causal explanatory questions. Resource allocators and senior decision makers responsible for budget setting tend to focus on causal descriptive questions. In practice, the distinction between causal explanation and causal description can be a blurred one. In some circumstances CIEs can provide an explanation of why certain impacts were found, for example through exploring the impacts of interventions on important subgroups. However, it is essential to consider carefully the types of questions that stakeholders have regarding an intervention, and to select the appropriate approach to answer each of them. In cases where the primary question is whether an intervention works, a counterfactual impact evaluation is in many circumstances appropriate. In cases where the primary question is how an intervention works, theory-based and process evaluation methods are more suitable.

These different levels of questions and purposes are summarised in the figure below.

This discussion leads to the conclusion that CIEs need to be developed within the evaluation plan. This evaluation plan has to comprise different forms of evaluation that are directed at answering different questions, for different policy stakeholders. In practice, an evaluation plan will seldom, if ever, incorporate a CIE without a process evaluation.

A wide range of approaches are deployed in the name of evaluation, and serve a range of different purposes. The critical question is how these approaches can be combined in useful ways to promote policy learning. Combining different types of evaluation in the appropriate way - with different purposes within the programming cycle - is the real challenge in this field. As has been discussed and as is shown in the next figure, CIE, process evaluation and theory-based approaches complement each other.

Figure 1 Different tasks and types of evaluation



Source: adapted from Martini A., 2009.

Combining types of evaluations

A solid evaluation strategy should comprise the following elements:

- Theory-based impact evaluation,
- Process (or implementation) evaluation,
- Counterfactual impact evaluation (CIE),
- Efficiency analysis.

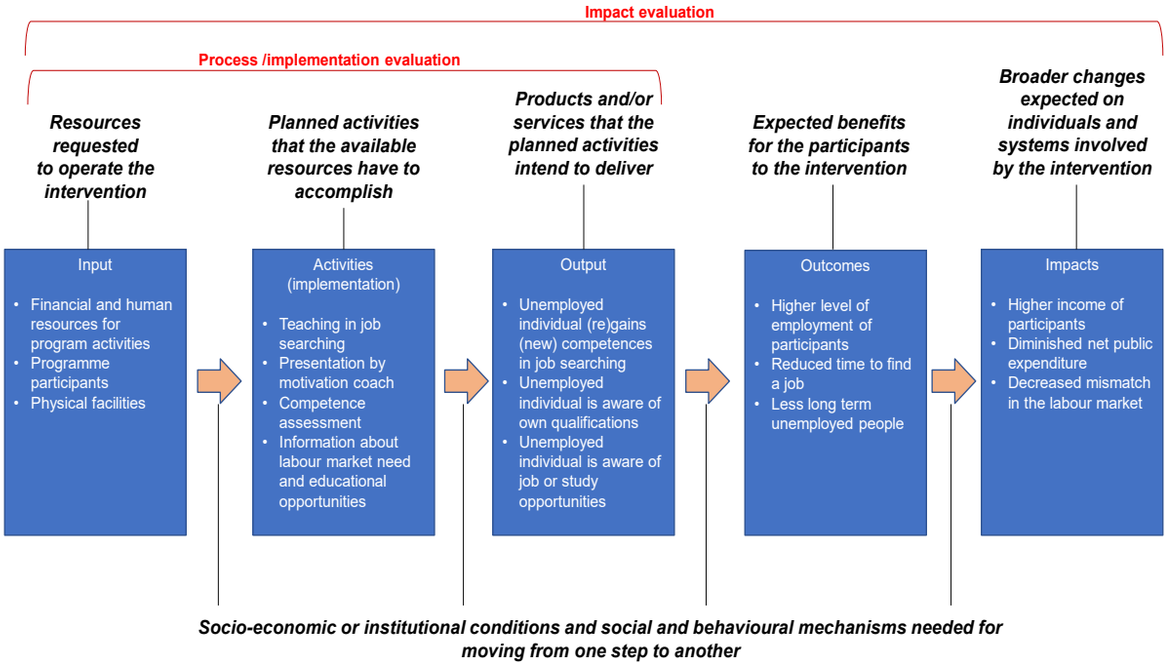
In this Guide, only counterfactual approaches to impact evaluation will be discussed in-depth. In the context of CIE, theory-based approaches are means of understanding the design intent behind an intervention.

Theory based evaluation refers to a 'theory of change'

Theory-based evaluations are used in some circumstances to describe the intended operation of the intervention and to test whether the change in results predicted by the intervention theory can be observed. In this sense, theory-based approaches can be used to assess impact in answering the question "how" an impact has been produced, and may be used to examine an intervention's impact where CIEs are not possible. The next figure shows a stylized example of theory of change for an intervention to improve employment services. A theory-based evaluation examines whether evidence confirms the sequence of causal steps, from input to impact, as suggested by the theory of change, and under what conditions and through which social mechanisms this was possible. A detailed account of the use of theory-based approaches to determining impact is beyond the scope of this document.

In the context of CIE, theory-based evaluation considers the way an intervention is planned and designed and how it is intended to operate. Essentially, the approach involves working with the stakeholders of an intervention on developing a shared account of an intervention's underlying 'theory of change', as simplified in the next figure. All interventions embody a programme logic which links inputs and activities to outputs, intermediate, and then longer-term results. Consequently, the articulation of a theory of change is an important support to CIEs too; it facilitates the reconstruction of the implementation process and its possible influence on outputs and outcomes, as well as the identification of the most representative outcome variables to be checked in the counterfactual analysis.

Figure 2 Illustration of the logic model approach or 'theory of change'



Source: Adapted from the W.K. Kellog Foundation, 2004 and Bredgaard T., 2015.

... adding to CIEs

Theory-based evaluation can link with counterfactual impact evaluations in a number of useful ways. A clearly articulated theory of change (or intervention logic) can inform the design of a CIE. Among other aspects, a well-defined theory of change can tell the designer of an impact evaluation the following:

- Which results are important and require measuring?
- What might be the likely sign and size of intervention impacts?
- Who is the intended target group and how can a control group be selected?
- How long might it take for programme effects and results to materialise?
- What data might be required in order to measure participation in the intervention?
- How plausible is the control group as a measure of the counterfactual?

Process evaluation

Developing a theory of change can also help identify potential unanticipated effects which can be taken into account in designing a CIE. To some extent, a clearly articulated theory of change may also help the evaluator interpret results from a CIE study. However, in terms of interpretation, a **process evaluation** can also be very informative.

Process evaluation in the context of CIE has two objectives. The first is to assess "fidelity", the other is to assess the difference between the experiences of the treatment and control participants.

Fidelity assessment

The "fidelity" assessment examines the extent to which an intervention - as delivered - is faithful to its design. A process evaluation considers what services were actually made available to intervention participants. Do they correspond to what is intended by the theory of the intervention? What

accounts for variation in delivery across sites, if variation is observed? Most interventions have both a management and effect logic:

- The management logic concerns how implementing bodies are expected to respond to programme rules and incentives.
- The effect logic concerns how the people who are targets of the intervention are expected to respond, given what is actually delivered.

The fidelity side of process analysis provides information on what was actually accomplished in an intervention and, therefore, what actually contributes to the observed effects. It also provides important feedback for project management.

Difference between treated and control groups

The difference assessment is particularly important in the context of counterfactual evaluation. It is common to focus, as has been done for much of this Guide, on intervention impacts. But before impact on results comes impact on inputs, the difference in opportunities between treatment and control groups that an intervention actually achieves. In principle, every CIE can be 'turned on its head' and the treatment group used as control for assessing the result for people in what was, before the inversion, called the control group. The implication is that as much needs to be known about what controls experience as is known about the treatment, because it is to the difference between treatment and control in inputs that CIE assigns causality for differences in results.

Returning again to the training scheme, one can imagine two quite different initial circumstances. In one, the training scheme is provided in a general context where nothing of the sort is otherwise available. The controls simply do without. But another possibility is that there are some substitutes. Training may be available, for example, from firms specialising in vocational preparation. If this is the case, process analysis needs to include, to the extent possible, assessment of the difference in training take-up between treatment and control, not just presume that all dimensions of the treatment are beyond reach of the control group.

CIE needs a process evaluation

While process evaluations can be commissioned completely independently of other forms of evaluation, their importance for both management and CIE makes it essential that process and impact evaluation be planned together.

Good process analysis can contribute to achieving fidelity, and process evaluations provide a causal explanatory account of an intervention. Without a well-designed process evaluation, it is often difficult to fully interpret the results from a CIE or to gauge the costs required for benefit-cost assessment, once impact estimates are at hand.

As noted above, one further contribution process evaluation can make to the interpretation of findings from impact evaluations, is an account of the context in which an intervention operated. Understanding context is important because it provides the conditions of success of the intervention, and a sense of the extent to which it might produce similar effects if implemented elsewhere, within different geographical areas, or at different points in time. This is especially important for discussing transferability of policy approaches and highlighting good practice in transnational learning and exchange. Process analysis contributes to confidence in what is termed the **external validity** of evaluation results.

Determining cost-effectiveness ratios

In most applications, efficiency analysis involves either an assessment of cost-effectiveness or a full cost-benefit analysis.

Cost-effectiveness analysis involves comparing the costs of the intervention to its effects or impacts that have been determined by a CIE. Put simply, a cost-effectiveness ratio is derived by dividing an intervention's impact - expressed either in the units of measurement or standardised units- by the net cost of delivering the intervention per treated unit.

CBA for comparing benefit with net cost

A cost-effectiveness ratio for a training programme that aims to help unemployed persons find work might reveal the funds needed per participant in order to move that participant from unemployment into work.

A cost-effectiveness ratio is an important measure for those responsible for allocating resources across programmes. Ratios obtained from a range of different interventions enable resource allocators to make relative judgements as to which interventions provide greater value for money.

Instead of expressing programme effects in either their unit of measurement or standardised units, a **cost-benefit analysis** (CBA) attempts to monetise the impact estimates obtained from a CIE and compare these to an intervention's net costs. The purpose of cost-benefit analysis is to determine whether the monetised benefits of a programme exceed its net costs. A cost-benefit analysis of a typical ESF training programme would compare the intervention's benefits for its participants, the government and society more broadly, to the net costs of the intervention. For participants, the benefits of the programme (usually improved employability and increased earnings) are obtained from a CIE. Subtracted from this will be the value of the taxes paid by participants and other costs of employment in order to obtain a net benefit. From the government's perspective, the benefits of the intervention will flow from additional tax revenues and reduced welfare payments, whilst the government would bear most of the costs of the intervention. The costs for society as a whole are derived from summing the benefits to participants and government and subtracting from these the sum of the costs to participants and government.

Cost-effectiveness analyses and CBA are still not very widespread among ESF evaluations. However, these analyses are very useful in deciding whether an intervention should be funded again in the future or in identifying the most effective intervention among a set of similar interventions (see the example in the following box).

Impact estimates from a CIE are a key ingredient in both cost-effectiveness and cost-benefit analyses. In the former, they provide the measures of effectiveness, while for the latter they provide a key source for estimating monetised benefits. What is also clear is that both cost-effectiveness studies and cost-benefit analyses require the collection of accurate cost data from which net costs might be derived. Such activities are usually referred to as a cost study. In some complex mixed-method evaluations, cost studies are frequently integrated into process evaluation, in which research instruments can be adapted in order to collect important cost data.

Box 1 An example of CIE for comparing benefits and costs

An example of the use of a counterfactual approach to estimate benefits and costs can be found in Bazzoli M. et al., 2018. The study focuses on vocational training programmes carried out in the Italian Autonomous Province of Trento in 2010-2011, providing more than 300 hours of training activities. Two groups of courses were evaluated: those financed by provincial resources (PVr) and those financed by the European Social Fund (ESF), involving 954 and 205 participants respectively.

The main steps for the implementation of CBA were the following: 1) the authors assessed the impact of training courses on the probability of participants finding a job during the three years after the start of their course, applying a propensity score matching[12]; 2) the impact of the courses on gross earnings up to the end of 2013 was estimated; 3) the authors estimated the amount of additional fiscal returns (deriving from effects on earnings) and the savings in public money generated by the reduction in the number of recipients of unemployment benefits; 4) the benefits and the costs of the courses were compared.

Several administrative datasets were used: a) monitoring data relating to participants and their characteristics; b) data from PES (Centri per l'impiego or Job Centres) registers of the unemployed to identify the control group; c) data from the *COB* database, the archive of companies' mandatory notification of labour contracts sent to the Public Employment Services and used to identify the employment status of treated and control groups, both before and after participation in the training course; d) data from tax revenue archives, to calculate the earnings of individuals and e) data from the Italian Social Security Institute (INPS) for information on unemployment benefits received by the individuals.

After 36 months the probability of being employed among the treated group in the PVr courses was around 5 pp higher than in the control group, while the impact of the ESF courses was much higher, about 28 pp. In the 3 years after the intervention, people who participated in the PVr courses earned on average a total of Eur 2,250 p.a. more than the control group, while people participating in the ESF courses earned Eur 4,106 p.a. more than the control group. Data also enabled the authors to estimate the benefits to public administrations in terms of increased tax revenues and decreased expenditure on welfare benefits. People in the PVr course group paid Eur 126 of income tax more than the control group for each year under consideration, while the estimate for people in the ESF course group amounted to Eur 318 p.a. In terms of decreased unemployment benefits (UBs) paid by the public administration, the impact of the courses was negligible, most probably because the monetary value of the UBs depended on the duration of employment prior to becoming unemployed, and many participants were young with scant work experience.

The costs of the interventions amounted to Eur 4,800 per participant in PVr courses and Eur 14,500 in ESF courses. The authors also compared costs and benefits at the individual level for both types of course; on average, it was found that costs were higher than benefits when considering the 2010-2013 period[13].

[12] More specifically, the authors applied the blocking with regression adjustment estimator.
[13] Another "similar" exercise can be found in Lammers M. and Kok L., 2021.

Chapter 2. Practical considerations in developing a CIE

Preparing a CIE

This chapter examines practical issues to consider when preparing for an evaluation. It should be used when planning evaluation activities, when deciding which interventions to subject to a CIE approach and to identify the key questions to be addressed when designing a CIE.

The starting position is assumed to be one in which a programme manager within an MA (or a manager of an intermediate body (IB) responsible for implementing an ESF intervention) is considering which interventions to evaluate, and the appropriate strategies for incorporating a CIE. It is also assumed that officials within an MA will not conduct evaluations themselves, but instead contract out or commission evaluation services from external experts. Although the CIE will be undertaken by a contractor, the MA (or IB) will have to plan and prepare for an impact evaluation prior to commissioning.

The evaluation strategy, including the various types of evaluations as described in the previous chapter, needs to be set out in the evaluation plan.

Evaluation plans are compulsory for all the programmes and must be approved by the Monitoring Committee no later than one year after approval of the programme[14]. The plans must be drawn up at the beginning of the programming period and include arrangements for the evaluation process (the governance of the evaluations and the link between evaluation and monitoring), actual evaluation activities (e.g., an indicative list of evaluations to be carried out, scope of each evaluation, main questions, necessary data, potential use, indicative timetable, management structure), timing of the evaluations, overall budget, and evaluation capacity building.

Evaluation plans tend to be general in nature, whereas planning a CIE requires more detailed preparation. Ideally, this should take place when the evaluation plan is drawn up, some details may follow at a later stage. However, MA/IB need to be aware that establishing the stakeholder connections and other arrangements necessary for intervention-related data collection is rarely easy and needs planning well in advance.

Developing an evaluation scheme for specific interventions

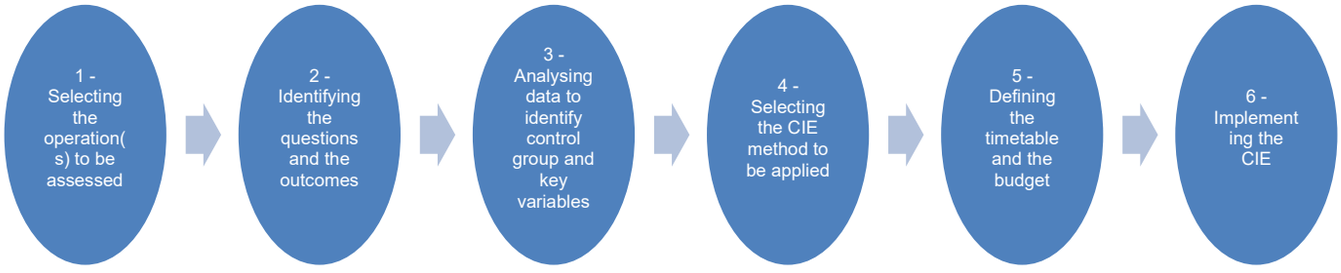
This Guide focuses on ways to develop an evaluation scheme for specific interventions that are candidates for CIE. This scheme should be part of the evaluation plan or, alternatively, might be established as an operational step following on from an evaluation plan.

Not all ESF-funded interventions can be the subject of counterfactual evaluation. Policymakers need to choose where to focus their attention. A process of selecting interventions for impact evaluation will be necessary. This Guide pinpoints some aspects MAs will need to take into account when selecting appropriate interventions. Furthermore, the central purpose of this Guide is to help those responsible for commissioning CIEs think through some of the challenges in constructing a successful impact evaluation, and in so doing, develop evaluation schemes for the various CIEs they are considering.

[14]Evaluation plans are required according to art. 44(5 and 6) of Council Regulation (EC) No1060/2021.One plan can include the planned evaluations of more than one programme, but all programmes have to be covered by an evaluation plan.

This Guide assumes that, after selecting the interventions for CIE, MAs will need to draw up an evaluation scheme for each chosen intervention. Here, the term 'scheme' is used to distinguish this activity from the formal evaluation 'plans' required through Common Provisions Regulation 2021/1060 for the 2021-2027 programming period. In particular, the term "scheme" refers to the set of standardized activities needed to define and implement a CIE, which have to be prepared in advance of its launch, as shown in the following figure.

Figure 3 Main sequence of activities of a CIE



Questions to be confronted in evaluation planning

These schemes will form the basis for commissioning CIEs and lay the groundwork enabling contractors to undertake a rigorous and valuable study. The remainder of the chapter looks at those questions which need to be confronted in evaluation planning. Evaluation schemes must be tailored to the specific circumstances under which the intervention operates. It is impossible to speculate as to what these specific circumstances will be. As a result, this Guide discusses questions that a) should be addressed in schemes, or b) should stimulate thinking about challenges that schemes will need to address.

Having reviewed some of the issues that need to be addressed when considering which interventions might be subject to a CIE, and whether it is even possible to undertake a CIE with the available types of data, attention now turns to some of the key questions that need to be considered in developing an evaluation scheme. This needs to be written before commissioning a CIE - or a wider evaluation study - in order to be able to prepare terms of reference and to appoint a contractor. The main content of such an evaluation scheme is listed in the table below.

Table 1 Recommended content of an evaluation scheme

Main steps in preparing and implementing a CIE	Content
1. Selecting the operation(s) for assessment	– The ESF operations, or types of operation (if possible, part of a pre-specified typology), to be evaluated with the CIE. – Summary description of the working logic of the selected operations (objectives, main eligibility criteria and target population, types of assistance, model of implementation, approximate dates of activation and completion, indicative budget).
2. Identifying the evaluation questions and the outcome variables	– evaluation questions – expected functioning of the operations ('theory of change') and outcome variables (e.g., employment status, changes in earnings, poverty status, average score in exams, etc.).

Main steps in preparing and implementing a CIE	Content
3. Analysing data to identify the control group and to measure the outcome variables	– administrative data (e.g., unemployment registers, tax registers, insurance administrative data, register of schools or database of students …) or other data (e.g., surveys, big data) to be used - variables of the database, or survey, to be used in the CIE - time series of the needed variables - main rules and issues for data access (direct accessibility by the MA, need for agreement with other administrations, privacy rules and constraints).
4. Selecting the CIE method	– Possible CIE method to be adopted (this can be detailed later on, but the use of randomised control trials or other methods should be indicated in advance to foster a consistent evaluation process)
5. Defining the timetable and the budget	– timetable of the CIEs, main milestones: a) decision on experimental or quasi-experimental approach, b) detailed evaluation questions, c) preparation of ToR, d) identification of the evaluator, e) data preparation, f) data analysis, g) intermediate and final reports, h) validation of the results, i) dissemination of results and lessons. – The budget available for the CIE
6. Implementing the evaluation	– Selection of the evaluator – Supervision of the implementation of the CIE – Reporting of the CIE – Distribution of results (main stakeholders to involve, main tools)

Box 2 CIE evaluation embedded in a wider framework

Many CIEs of ESF-financed interventions conducted across Member States are embedded within wider evaluation frameworks:

- in Germany, under the OP Bund ESF 2014-2020, the counterfactual evaluation of the programme for integrating long-term unemployed into the labour market funded under the IP 9.i is part of a broader evaluation strategy envisaging annual interim reports which also examine themes related to the implementation of the interventions. From 2017 to 2021 four reports were produced[15]. The same applies to the evaluation of the ESF measures supporting the integration of long-term unemployed in Baden-Württemberg, where the counterfactual analysis followed a more qualitative analysis focused on participants' assessments of the interventions[16].

- In Piedmont (Italy) the counterfactual evaluation assessing the employment effects of employment vouchers for vulnerable people financed under the IPs 8.i and 9.i was one step in a more comprehensive ongoing evaluation. Two initial reports examined implementation issues and the subjective perceptions of the participants (April 2018 and February 2019), while two further reports in July 2019 and at the end of 2020 focused on employment impacts using a counterfactual approach[17].

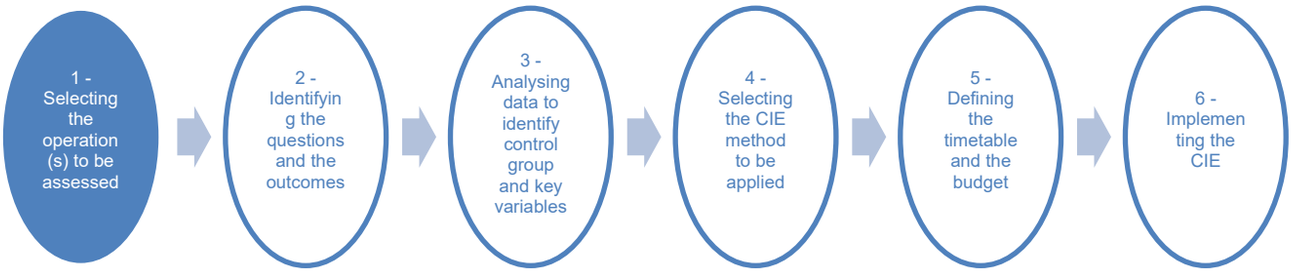
- In Marche (Italy), the CIE examining the impacts of interventions for long-term unemployed in 2020 is an impact analysis with a thematic focus, following a more general 2019 impact analysis of ESF operations aimed at the unemployed (Placement Report) [18].

- In Poland, the OP Knowledge Education Growth programme 2014-2020 commissioned a number of evaluations (8 reports[19]) between the end of 2015 and May 2020 to analyse the ESF and YEI support

[15] See Boockmann B. et al., 2017 - Boockmann B. et al., 2018 - Boockmann B. et al., 2019 and Boockmann B. et al., 2021. In 2019 and 2021 CIE was applied
[16] See Hunger K. and Sattler K., 2017 and Scheller F. and Seidel K., 2020. In 2020 CIE was applied.
[17] See Pomatto G., 2017 – Pomatto G., 2019 – Poy S., 2019; Poy S, 2020. In 2019 and 2020 CIE was applied
[18] See Pompili M., Giorgetti I., 2020 - Pompili M., Giorgetti I., 2020a
[19] See Instytut Badań Strukturalnych - Imapp - IQS, 2015 – Baran J. et al., 2016 – Baran J. et al., 2017- Baran J. et al., 2018 - Baran J. et al., 2018a – Palczyńska M. et al., 2019 - Kalinowski H., 2020 - Kalinowski H. et al., 2020. In 2017 and 2020 CIE was applied.

for young people from different perspectives and applying various analytical methods (e.g., qualitative analyses through surveys and interviews, macro models, field activities, and a counterfactual approach).

2.1. Selecting interventions for impact evaluation



Criteria for selecting interventions

Selecting interventions for impact evaluation requires three key steps:

1. Strategic issues must be identified;
2. Once strategic priorities are clear, individual interventions must be assessed as to whether they conform to the basic requirements of a counterfactual approach, and to what extent they are innovative and/or make a significant contribution to the knowledge base.
3. The availability or potential availability of the types of data required to conduct a CIE must be made clear. This third issue has hitherto proven to be a major barrier to conducting counterfactual evaluations of ESF interventions and deserves particular attention.

Box 3 Questions for selecting interventions for a CIE

CIE is not appropriate for all interventions and conducting CIEs for all of them is generally not cost effective. Managing authorities have to decide how to allocate resources so as to achieve the greatest benefit. The evaluation plan is expected to reflect these choices and in planning CIEs three main elements should be considered: a) strategic priorities, b) the feasibility of a CIE, and c) availability of necessary data.

The evaluation strategy is influenced by scale, policy development, areas of uncertainty and the need for knowledge. MAs should ask the following questions:

- Do the large amounts of funds allocated to this intervention render it particularly important to justify expenditure? These interventions are relatively easy to identify because they receive the bulk of funds allocated to each specific objective (as defined in art.2 of the ESF+ Regulation 2021/1057).
- Is the measure the focus of a reform process and are results from the evaluation likely to contribute to a critical review of the effort? These interventions are linked to recent reforms of labour, education or social policy; while they may not receive large amounts of money, they are nevertheless crucial for the success of the reform.
- Is the intervention innovative and being tested through a pilot or trial before being scaled-up? These interventions might not receive huge resources, but require early assessment to decide whether to continue and expand, or end the experiment.
- Does the intervention focus on areas which require additional evidence of effectiveness? This group includes interventions which have not been evaluated in the past, or whose last evaluations are so old as to require an update. In the ESF+ programmes these interventions are numerous because CIEs have only recently become widespread and 'net effects' are unknown.

Feasibility relates both to the characteristics of interventions and to the circumstances in which they are

introduced. Planners should be able to give affirmative answers to the following questions:

- Is the intervention treatment discrete, distinctive and sufficiently homogenous?
- Is the comparison between treatment and control groups meaningful enough to measure impact?
- Is the target population large and well-defined?
- Is the theory that links the intervention to intended outcomes logically coherent?
- Can the treatment group be clearly identified within the target population?
- Is the size of the treatment group sufficient?
- Can a credible control group be identified?
- Can the difference between treatment and control experiences be maintained over a long enough period to gauge impact?

Data are critical. The essence of CIE is measurement, and measurement requires quantitative information, both on treatment and control groups and the context in which the evaluation is conducted. Exactly which data are required is usually determined by the theory of the intervention and the strategy employed for establishing the counterfactual. In selecting interventions for CIE, an MA planning a CIE needs to ask:

- What is it essential to know about members of the target and control groups?
- What is it essential to know about the nature of the intervention as actually delivered to the treatment group?
- Does the control group receive no or any other treatment? Are data available on this?
- What data are available from administrative and other sources?
- Are data available that describe individual careers?
- Can individualised data from various sources be linked?

More detail on these issues is provided later in this chapter.

2.1.1. Prioritising interventions for impact evaluation

Before prioritising specific interventions for CIE, wider strategic issues need to be considered. The focus should be on selecting those interventions for which impact evaluations promise the greatest return in learning about what works. The benefits stemming from well-designed, rigorous evaluations accrue not only to the MA and MS that commission them, but also to other MSs and their MAs, to other stakeholders, and to the Commission.

Contribution to justifying expenditures

Focus on resource-intensive interventions

Given the focus of CIEs on addressing questions that are critical for policymakers, particularly those responsible for resource allocation decisions, it makes sense to focus impact evaluation efforts on programmes and interventions that are particularly resource-intensive. The more time and other resources a particular programme or intervention absorbs, the more important it is to understand whether it works, and therefore whether the benefits generated exceed the costs incurred. Expensive interventions that do not produce social or economic value may need to be reconsidered, while others with evidence of added value may deserve increased funding and attention.

Results from recent evaluations of ESF interventions funded in the 2014-2020 programming period have shown that strategic adjustments and increasing concentration on key policy objectives are necessary. Employment and labour mobility interventions were shown to be less effective for older people and

those more distant from the labour market; these groups require new and more effective instruments[20]. Although social policy interventions show a large variance in cost per participant and type of operation, there is generally no systematic cost-benefit analysis[21]. Education and training interventions showed positive results but the limited number of impact studies prevents a complete assessment of their long-term effects[22]. CIEs offer the prospect of being able to sift interventions in order to identify the most effective and efficient approaches for given target groups, thereby maximising the 'value for money' of the new ESF+ programmes.

Interventions contributing to policy innovation

Contribution of an intervention to a reform process

Interventions that form a key component of a broader reform programme will often attract significant funding. The fact that an ESF intervention is central to a social inclusion strategy, or a critical feature of an active labour market programme, will naturally focus greater attention upon it.

Innovative and exploratory

New and innovative pilot interventions are obvious candidates for CIE. Testing the effects of interventions through a pilot or trial clearly requires rigorous evaluation. Evaluating through a well-designed CIE is all the more important where there is a clear commitment to scale up or roll out the intervention more widely should it be perceived as being successful.

Contribution to learning

Producing evidence

The case for focusing attention and resources on specific programme areas - and specific interventions within these areas - is reinforced where there is little or no existing evidence regarding what works within the policy area concerned. For instance, where there is genuine uncertainty about policy in the future, and a risk of over-reliance on evidence that may not be directly relevant (e.g., evidence from other countries).

High quality evaluations can be considered a public good. The benefits they generate in terms of learning extend to stakeholders beyond those within a specific MA. As a result, it is important to consider which stakeholders might stand to benefit from the proposed impact evaluation. These may be intermediate bodies (IBs) or agencies dealing with interventions within the same programme, other MAs or IBs in the Member State concerned, or agencies and institutions dealing with national or regional funds. Another obvious external stakeholder that should be considered is the European Commission, and there are also stakeholders in other MSs who might learn from an evaluation. Taking into account the needs of those beyond the immediate stakeholders is an important contribution that policy makers and programme managers can make to mutual learning.

Championing CIE methods

A final strategic consideration in selecting areas for attention when developing CIEs, is to consider those interventions that can showcase the benefits of CIEs and act as a model.

2.1.2. Selecting interventions that are suitable for a counterfactual approach

[20] Fondazione G. Brodolini, Metis GmbH, Applica, Ockham IPS (2020).
[21] ICF, Cambridge Econometrics and Eurocentre (2020).
[22] Ecorys, Ismeri Europa (2020).

Having considered wider strategic concerns that might motivate the selection of particular interventions for CIE, this section looks at the specific characteristics of interventions that might make them suitable for a counterfactual approach. Such characteristics are many and varied. Some features of an intervention might lend themselves to a CIE in one set of circumstances, but in another frustrate attempts to implement it. As a result, it is not possible to provide a comprehensive list of considerations. However, the features of interventions that are more likely to lead to a successful CIE are worth a mention.

Box 4 Most common types of interventions and target groups chosen for ESF CIEs

The majority of CIEs of ESF interventions are focused on active labour market policies directed towards the unemployed and subgroups among them affected by a specific disadvantage. The large number of impact analyses of interventions for young people, reflects the regulations for the YEI requiring such evaluations to be carried out at specific intervals.

About half of the CIEs identified in the Evaluation Helpdesk[23] project since 2015 relate to thematic objective 8 “promoting sustainable and quality employment and supporting labour mobility”. Similarly, in most CIEs examining interventions financed under TO 9, attention is on the effectiveness of the interventions in integrating vulnerable unemployed into the labour market (for example the German intervention targeted at long-term unemployed (LTU) financed under the OP Bund ESF 2014-2020).

The most commonly analysed forms of support provided to the unemployed are training, internships or other forms of work experience and subsidized jobs. This emerges clearly from the examples analysed in this Guide (see the examples of the Italian evaluation in Marche, the evaluation training courses for migrants in Germany, the vocational training for NEETs in Latvia). The analysis of counselling and job matching services is less common (as in the case of the Swedish example in this Guide, where a pilot action for intensifying support to the unemployed for PESs is evaluated through a randomized approach). ESF interventions supporting self-employment or business creation are not often evaluated, according to Helpdesk data[24].

In some cases, the CIEs analyse different types of interventions in a pooled way, incurring the risk of mixing different intervention logics, and reducing the reliability of the comparison between treatment and control groups.

The attempt to evaluate the effects of the interventions for vulnerable people is noteworthy, not only in terms of employment results, but also in terms of “soft outcomes”: a CIE conducted in Germany focused on this aspect, by assessing the impact of job creation schemes for LTU on perceived health measures, life satisfaction, sense of belonging and social status indicators.

ESF interventions in the education area are less frequently assessed through a counterfactual approach. Fewer CIEs relate to thematic objective 10 than to TO 8 or 9, and the focus is often on interventions financed under IP 10.IV, concerning vocational education and its effects in terms of integration into the labour market. Two reasons for this are more limited accessibility to the datasets and more stringent privacy rules[25]. Nevertheless, attempts in this direction have been made: in Spain (Asturia) an intervention in secondary schools to discourage early school-leaving was assessed; in Poland (Podalskie) the CIEs examined the effects of a project aimed at promoting vocational education among young students; in Portugal, grants supporting students in higher education were analysed.

CIEs can be conducted at various policy levels (i.e., one or several Priority Axes, specific objectives or operations[26] in a Programme), and may cover national or regional ESF+ programmes, may focus on homogenous target

[23] See https://ec.europa.eu/regional_policy/en/policy/evaluations/member-states/
[24] Some examples are the following: Borik V. et al., 2015 - Ires Piemonte, 2019 – Openfield, 2019.
[25] See for example Ismeri Europa – Ecorys – Institute for Employment Studies, 2019.
[26] According to art.2(4) of the Council Reg. (EC) No1060/2021: ‘operation’ means: (a) a project, contract, action or group of projects selected under the programmes concerned; (b) in the context of financial instruments, a programme contribution to a financial instrument and the subsequent financial support provided to final recipients by that financial instrument.

groups (male or female, youth or vulnerable populations, long-term unemployed, etc.) or types of intervention (e.g., training, services for social inclusion, or discouraging early school-leaving (see previous box)).

Examples from Member States indicate that a variety of instruments used within ESF+ are appropriate for CIE, including training, employment incentives and labour market services (e.g., job counselling, coaching). On the other hand, job rotation and job-sharing interventions, start-up incentives or support for systems and structures, as well as interventions in the fields of education and social inclusion, are more challenging to conducting a CIE.

It is instructive to consider which interventions are more promising from a CIE perspective by considering the following questions:

Is the intervention discrete, distinctive and relatively homogenous?

Clearly distinguishable treatment

The treatment or treatments delivered by an intervention need to be distinguishable from those in other interventions. Moreover, there needs to be a meaningful contrast between what an intervention's participants receive and what other similar groups of individuals benefit from. If treatments are blurred to the point that it is not possible to identify a discrete group of recipients, then counterfactual approaches are not possible or desirable.

CIE methods become very complex and difficult if the treatment status of a given unit (an enterprise or individual) affects the potential result of other units (through so-called wider 'general equilibrium effects'). In training programmes, this can occur when graduates from the programme make it difficult for other non-trainees to find work in the short run. Where this is thought to be a substantial problem (for example in the case of large-scale interventions), macroeconomic analysis may be required to assess the extent of substitution and displacement effects. MAs should obtain expert advice where such effects are likely to be present.

Homogeneous interventions

The intervention itself should be relatively homogenous. This means that all participants in an intervention should receive or be exposed to broadly the same package of measures. There are a number of implications for CIE if the range of measures delivered to participants within a single intervention is too diverse. First, it might not in reality make sense to talk of a coherent intervention, but rather interventions with separate causal processes at work; second, it will be difficult to interpret impacts that are reported as average net effects over a group of disparate interventions; third, subgroup analysis might be warranted but, if there are too many subgroups within a treatment group, sample size limitations may constrain the ability to report usable findings.

Is the treatment being compared to no treatment or do other relevant forms of treatment exist?

ESF is co-financing national and regional labour market and social inclusion policies. Thus, any CIE evaluation scheme needs to take into account carefully whether the intervention is clearly identifiable and individuals have the opportunity to receive services from other (national or regional) programmes and funding sources. What is important is that the treatments being evaluated actually alter the opportunities or resources available to participants compared to what is available to controls and that differences can be measured and monitored.

Complex treatments

Such 'complex treatment' issues tend to be context-specific. They complicate CIE design and implementation. Their presence underscores the importance of careful evaluation planning - developing the evaluation scheme - in advance of implementation.

Is there a large and well-defined target group?

Large sample size

CIEs require large sample sizes relative to some other forms of evaluation. Target groups composed of individuals in adequate numbers are essential, and it must be possible to locate control groups of sufficient size. This issue is discussed in more detail below.

Establishing the identity of the target group

It is important that the intervention being considered for CIE is targeted at a well-defined group. Without a clear understanding of who the target groups are, it is difficult to identify a meaningful control group. Some interventions deliberately seek to recruit individuals into treatment through informal mechanisms, encouraging processes that are not predefined or too prescriptive (e.g., difficult social targets, such as young NEETs or disadvantaged groups, can be involved through occasional and case-by-case procedures); this can make it difficult to identify precisely the treated individuals and the related control group.

Is there a clear causal mechanism?

Distinct policy mechanism

As mentioned previously, when main evaluation methods and 'theory of change' were presented, it is often useful for a theory-based evaluation to have been conducted in advance of or in combination with a CIE. Developing a theory of change, or a detailed logic of the intervention, can help those designing a CIE in a number of ways; most importantly, in determining whether an intervention has a coherent causal mechanism which underpins it. Interventions without a clear and convincing causal mechanism are unlikely to produce impacts of sufficient magnitude to be identified statistically through a CIE.

Can outcomes be defined quantitatively?

Need to measure results

There is a need to obtain quantifiable measures of outcomes (or results). Such data and indicators may be obtained from administrative sources, or specifically targeted surveys.

In some circumstances, interventions may have intended results that need specific provisions to be measured quantitatively. For example, an intervention might be concerned with changing attitudes, beliefs or opinions. In such cases, surveys need to be administered to measure these changes. Some interventions have quite vague or poorly defined results. Again, the development of an intervention logic can help sharpen understanding of what an intervention is seeking to achieve and how it intends to bring about change in the results of interest.

Is the intervention introduced in a way which makes it possible to find a meaningful control group?

Selection mechanism for treatment

In order to identify a meaningful control group, it is important to consider how treated units (persons or enterprises) are selected for an intervention or why they decide to take part; whether the same data source – e.g., the same survey instrument - can be administered to the control sample as to the treatment group; and finally, whether it is necessary to select control samples

who are subject to the same labour market conditions as the treatment group. Some examples are highlighted in the box below.

If an intervention is mandatory and delivered to the entire target population more or less simultaneously, it might prove difficult to locate an untreated portion of the target population to act as a control.

Box 5 Defining control groups

In the examples of CIEs examined in the Guide, the selection of control groups was driven by the characteristics of the interventions (for example eligibility criteria) and also by the availability of appropriate data.

In comparison with the experience in CIEs in the previous programming period, the identification and selection of the control group is more often based on administrative data, especially the unemployment registers. The most common strategy is to identify potential control individuals with similar characteristics, registered as unemployed at PESs during a certain period, as required by the eligibility criteria. This applies to some evaluations in Italy (Marche, Province of Trento and Piedmont), Poland (Lubelskie and Podlaksie), as well as the German evaluations of training courses for migrants, job creation schemes, and integration measures for the unemployed in Baden-Württemberg.

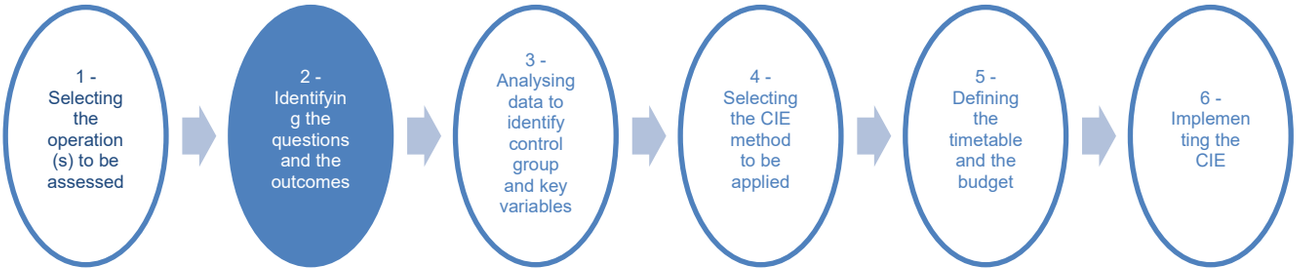
In other evaluations, though not covered by the examples presented in detail in the Guide, people who had enrolled but were not selected for treatment were subsequently selected to build the control group. This is the case in the Italian evaluation of Youth Guarantee, the German evaluation of the pilot project 'Citizen labour funded in 2007-2013', and the evaluation of the PIPOL programme implemented in Friuli Venezia Giulia[27]. However, this strategy is not used frequently, since in most countries and regions the monitoring information systems do not include information on people who applied but did not participate.

The German approach to the evaluation of interventions for LTU[28] was different, since the evaluation assessed the "intention to treatment" (ITT) and not the "average treatment effects on treated" (ATT). In this case, the treated group was composed of people potentially eligible throughout the implementation period of the programme, regardless of whether or not they were actually treated (from 2015), and the control group was composed of people with the same eligibility criteria but who were LTU prior to the implementation of the programme (2010-2012).

Only in the Swedish example, which applied a randomised approach, was the control group identified randomly: by design, the treatment (intensified support by PESs) was provided randomly to a group of young people, while the control group received the ordinary support offered by PESs.

In the few evaluations focusing on ESF measures for enterprises, the demarcation line between treatment and control groups was drawn between funded and non-funded applicants as in some Danish evaluations, which compare the performances of enterprises funded through the ESF with those of a sample of companies with similar features but which did not receive the support[29].

2.2. Evaluation questions and the outcome variables



[27] See respectively Isfol, 2016 – IAW Institut für Angewandte Wirtschaftsforschung - ISG Institut für Sozialforschung und Gesellschaftspolitik, GmbH, 2015 - Ismeri Europa, 2018.
[28] See Boockmann B. et al., 2019.
[29] For example: Danmarks Statistik et al., 2017 and Danmarks Statistik et al., 2018.

2.2.1. What are the aims and objectives of the intervention?

Combining CIE design with insights from intervention logic

In setting out an evaluation scheme, it is first of all advisable to describe the aims, objectives and key features of the intervention.

In many cases, documents that set out the aims and objectives of the intervention will already exist. However, it is important in the case of a CIE to be specific about the results and changes the intervention wants to achieve and, therefore, the impacts that are expected.

It is often beneficial to articulate an intervention's theory of change which sets out the means by which its various inputs and activities are intended to link to outputs, outcomes (or results) and thereby impacts.

2.2.2. What is the purpose of the evaluation?

In developing an evaluation scheme for a CIE, it is important to consider carefully the purpose of the evaluation. Without a clear understanding of why the evaluation is needed, it is unlikely that it will produce the evidence required. In the context of evaluations of ESF financed interventions, a series of questions needs to be considered:

- What is the purpose and nature of the evaluation in the context of EC regulatory requirements and guidelines?
- Who are the evaluation's main stakeholders?
- What use will be made of the evaluation results?
- What specific questions will the evaluation need to address?

What are the aims and the nature of the evaluation?

The aim and the nature of the evaluation

Firstly, **the motivation** for carrying out the evaluation needs to be defined. According to Regulation 2021/1060, "the Member State or the managing authority shall carry out evaluations of the programmes related to one or more of the following criteria: effectiveness, efficiency, relevance, coherence and Union added value, with the aim to improve the quality of the design and implementation of programmes. Evaluations may also cover other relevant criteria, such as inclusiveness, non-discrimination and visibility, and may cover more than one programme."[30] As shown above, the findings of a CIE generally relate to effectiveness (to what extent the expected outcomes have been achieved) and efficiency (the cost-effectiveness or the cost-benefit ratio of the intervention).

More generally, the EC encourages Member States to follow result-oriented approaches in their policy-making and to carry out evaluations that meet internal MS demands in their scope, design and timeframe. In this respect, in the ESF+ programmes, CIEs can also be implemented to answer specific evaluation questions or in accordance with national evaluation policy.

Secondly, **the nature of the evaluation** needs to be established:[31]

- Evaluations of an **impact nature** examine the effects of a programme, or group of programmes, in relation to EU and national priorities (this may be

[30] Art. 44(1) 1060/2021 CPR.
[31] See European Commission, 2007.

the macro-economic impact of the ESF, a focus on specific policies and themes, or horizontal priorities like childhood and equal opportunities).

- Evaluations of a **process (or implementation) nature** support the implementation of a programme, analysing progress and implementation methods and providing recommendations on improvements to the programme.

In principle, the counterfactual approach can be applied to impact evaluations, while process evaluation requires other methods (see also Figure 1 above).

The CPR does not require a specific number of impact evaluations in the 2021-2027 programming period, in contrast to 2014-2020. Instead, it asks for an evaluation strategy capable of assessing how support from European funds has contributed to reaching the goals of the programme with regard to all the main strategic profiles.[32] In addition, the CPR does not dictate which priorities or interventions should be the focus of the evaluation, but leaves this decision to the MA and the evaluation plan. This means that the evaluation strategy of each individual programme has to define the mix of impact and process evaluations and on which priorities and interventions to concentrate its major efforts. A final and general impact evaluation has to be produced by June 2029[33], but no other restrictions constrain the timing of the other evaluations, which are to be set out in the evaluation plan.

Who is the main audience?

Identifying the stakeholders

The evaluation's audience should be defined. Depending on the nature of the evaluation, it might include policy makers, MAs and programme managers, other MAs or implementing bodies in the Member State, or national or regional authorities running similar programmes. Where data are provided by institutions outside the programme management, these bodies ("data owners") should also be considered stakeholders. It is important to include all major stakeholders in an evaluation steering group in order to establish joint ownership of the process of designing and conducting the evaluation, as well as some experts in evaluation coming from academic or public institutions, to provide some technical advice to the MA.

What use will be made of the evaluation results?

Once the audience for the evaluation has been identified, use of the findings can be determined. Practically, this can be achieved through involving the steering group in the development of the evaluation questions, and discussions on the terms of reference.

Two key decisions to which CIE results frequently contribute are:

- Whether an existing intervention should continue, and
- Whether a new type of intervention should be implemented more widely (i.e., scaled-up).

In the first case, a CIE may attempt to assess the effectiveness of an existing or ongoing programme where budgets are under pressure, and there are potential alternative uses for the resources involved. In this situation, it is unlikely that the intervention has been evaluated previously using counterfactuals.

[32] Art 44 (1) of 1060/2021 CPR.
[33] Art 44 (2) of 1060/2021 CPR.

In the second case, interventions might have implementation constraints. For example, it may be implemented in a particular region or area of an MS, or for only a limited time period. In these contexts, results from a CIE may be used to determine whether the intervention concerned is effective and can therefore be usefully implemented elsewhere. Interventions in such situations are referred to as being piloted, or tested before wider rollout.

What questions need to be answered?

Once the intervention's objectives and the evaluation's purpose and ultimate uses are established, and the audience is clearly identified, it should be possible to specify in some detail the questions the CIE will need to address. In many circumstances, a range of audiences and stakeholders will have questions of a causal nature they will want the CIE to explore. The MA or the evaluator should collect these questions through extensive consultation, taking into consideration different points of view and suggestions from people involved at different stages of the intervention. Subsequently, the MA and/or evaluator have to prioritise the questions and focus the CIE on those more important and appropriate.

Key research questions

Some of the issues that might be considered in finalising a list of key research questions for a CIE include:

- Did the intervention produce or contribute to the intended outcomes in the short, medium and long term? And, did short-term effects differ significantly from those in the long run? Questions addressing these issues should be prioritised.
- Is it possible to have a quantitative measure of the outcomes? CIE has to rely on an adequate set of data (administrative data or direct survey sent to participants) independent of a preferred source.
- To what extent can changes in the participants' circumstances, or in the socio-economic context, be attributed to the interventions? This type of question requires a measure of the net effects of the intervention for comparison to control participants and context indicators.
- Were the effects of the intervention the same for all members of the target group? For example, was the impact of an intervention targeted at the long-term unemployed the same on males and females? And on those under 25 years of age, or over 50? The CIE capacity to examine effects in different sub-groups is powerful if the number of individuals in the treatment and control groups is sufficiently high.
- Has the intervention been cost-effective (compared to alternatives)? And what is its ratio of cost to benefit? This analysis of efficiency requires a CIE to produce an accurate measure of the effects, as well as precise information on direct and indirect costs and benefits.
- Is an ample amount of information and knowledge available on the impacts of similar interventions? To what extent is this knowledge applicable to the intervention under examination? When a policy is well-known and its effects have been extensively investigated, it may be useful to focus the evaluation questions on specific aspects of the policy, and so avoid repeating other analyses. A detailed screening of existing literature can inform decisions in this respect and it is, in any case, a useful support when designing the CIE.

It is important to have a clear idea of the range of research questions that a CIE will need to address prior to commissioning. Discussion of the questions the evaluation will address is a key element of any evaluation scheme.

Prioritising questions

It is important to prioritise questions and not succumb to the popular tendency to over-load an evaluation with too many questions. Ensuring that the evaluation is relevant to a range of stakeholders with differing interests, and making the evaluation tractable is a difficult balancing act. If an evaluation has address too wide a range of research questions, it can lose focus and end up addressing a wide range of concerns in a sub-optimal manner. It is often a case of 'less being more' - prioritisation is a critical phase in the evaluation planning process.

To prioritise evaluation questions, it is necessary to exclude duplications and assign a score to each question according to pertinent principles. These may include: the importance and real commitment of the stakeholder who formulated the question, the appropriate fit and congruence of the question with the programme's theory of change, the relevance of the question to the general purpose of the evaluation, the feasibility of the question in relation to available data, time and resources[34]. The resulting rank will order the questions by importance and allow selection of the most relevant ones.

In some cases, it is possible to nest a group of evaluation questions under a more general question; for instance, under a question like "what has been the net effect of the intervention". Other questions such as the effects on different groups of participants, perhaps even at different periods in time, could be nested. This nesting process, however, always has to give rise to a manageable and feasible set of questions.

What evaluation criteria can be associated with the evaluation?

Evaluation criteria and evaluation questions

The relation between evaluation criteria and questions has been mentioned above, but it deserves further clarification. Evaluation criteria (efficiency, effectiveness, EU added value, coherence, etc.) are necessary to assign a value to any collated evidence and reach an assessment of a policy (efficiency, etc.); while evaluation questions are necessary to make the demand of the commissioner explicit and focus on the main policy issues at stake. However, evaluation criteria and evaluation questions are connected. Every question is generally referable to a specific criterion and this relation is important for MAs and evaluators alike, because it links the evaluation design, necessarily based on the questions, with the requirements contained in the 2021-2027 CPR, which are related to the evaluation criteria in a different way.

Some examples of typical evaluation questions grouped with their proper criteria are presented below. They are selected and adapted from the 'Better Regulation Toolbox[35] (Tool #47 on *Evaluation criteria and questions)* of the European Commission to show examples consistent with questions that require an impact analysis and, where possible, a CIE:

- *Typical examples of effectiveness questions*
  - What have been the quantitative effects of the intervention?
  - To what extent can these changes/effects be credited to the intervention?

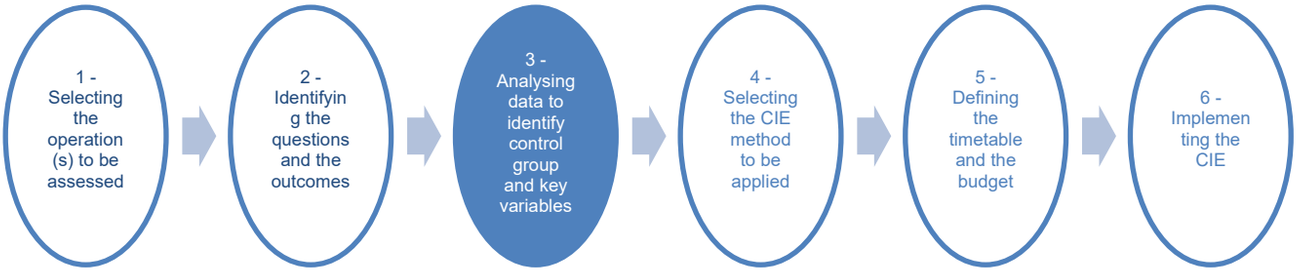
[34] See, for instance, Centre for Disease Control and Prevention CDC, 2013.
[35] See European Commission, 2017.

- To what extent can factors influencing the observed achievements be linked to the EU intervention?

– *Typical examples of efficiency questions*
  - To what extent has the intervention been cost-effective?
  - To what extent are the costs of the intervention justified, given the changes/effects it has achieved?
  - If there are significant differences in costs (or benefits) between territories, what is causing them? How do these differences link to the intervention?

– *Typical questions on EU added value*
  - What is the additional value resulting from the ESF+ intervention(s), compared to what is produced by similar national and/or regional interventions?
  - What would be the most likely consequences of stopping or withdrawing the existing ESF+ intervention?

2.3. Data to identify the control group and to measure the outcome variables



2.3.1. Are the appropriate data available or can they be made available?

Discussions held with MAs and evaluation experts from across the EU suggest that access to appropriate data is one of the key challenges in implementing CIEs; a key practical consideration is whether the types of data required are available. In this section, a simplified categorisation of the types of data required is presented, along with discussion of the sources from which such data might be obtained, or the types of primary data collection exercises that might be required. The crucial issue of data protection is also addressed.

Planning data collection

An important point concerning proper planning needs to be made. To a certain extent, attempts to implement CIEs have been thwarted in the past by a lack of data, because adequate plans were not put in place early enough. For existing interventions, it is important to identify members of treated and non-treated groups, and establish mechanisms to collect data from them, as they will be the focus of the evaluation. For new interventions, early steps should be taken, to ensure that the right types of data are collected at the appropriate times.

Main data sources

What types of data are required?

Broadly speaking, three types of data are required in order to conduct a CIE. In some instances, a single data source may contain one or more of these types. These are: treatment and control group records, outcome records, and context data records.

- **Treatment and control group records:** data sources are required which enable the evaluators to identify individual treatment and control group units (enterprises, persons or potentially geographical areas).
- **Outcome records:** as shown in Figures 6 and 7 in Chapter 3 of this Guide, CIEs require outcomes to be measured for both treatment and control groups. Ideally, data on outcomes for both groups should be gathered using the same collection methods and result measurements made at the same points in time.
- **Context data records:** data are required to enable the selection of well-matched control and treatment groups, and to allow any remaining differences between the two to be checked in analysis. It is important to collect as much data as possible on unit characteristics and factors which may be related both to the choice to participate in an intervention and to potential results, particularly result indicators measured pre-treatment. Context data might also include those which describe local labour markets (for example, local unemployment rates or measures of labour market tightness) and those which will enable analysis by subgroup.

Table 2 below sets out these three data types and suggests sources from which they might be collected. Examples of data used for ESF CIEs are given in the first box below, while the second box illustrates examples of integrated dataset used in analysis of labour market issues and in evaluations of labour market policies.

Table 2 Data types and sources

Data types	Sources
Treatment group records	- Intervention participation records (generally, held by beneficiaries) - ESF+ monitoring data (intervention characteristics, starts and completions, referral records, application records)
Control group records	- Administrative data such as social security, education and unemployment benefit records - Participation records (those who were eligible to participate but did not do so for reasons other than eligibility[36]-) - Existing national surveys, such as the LFS
Outcome records (required for both treatment and control groups)	- Administrative data: e.g., social security and unemployment records can also be used to put together outcome measures (benefit/social security receipts), national insurance and tax records (earnings, and employment outcomes) - Administrative records from education (standardised tests on achieved competences, graduation rates, enrolment and attendance) - Official company census or tax records (productivity or turnover before and after in-house training or new hirings) - Employment or output records from official statistics (in territorial counterfactual analyses to measure employment and GDP levels) - Bespoke surveys of treatment and control groups
Contextual data/ control variables (required for both treatment and control groups)	- Administrative systems – e.g., benefit records providing pre-treatment claim histories; national insurance and tax records, historic earnings and employment records - Official statistics on labour market or education (e.g., Labour Force Survey, basic data at regional or national level which also provide micro-data at individual level for specific elaboration) - Surveys of control and treatment groups. Where treatment rules are clear, control groups can be identified ex-ante and baseline data collected - Monitoring systems - in some circumstances, monitoring systems can be used to collect measures from both treatment and control groups, for example application systems where failed applicants can be used as controls.

Box 6 Examples of data used for CIEs

Among the examples examined in detail in this Guide, German evaluations are probably those relying on the richest administrative datasets. The Integrated Employment Biographies (IEB), integrate a number of different administrative sources of information managed by the Federal Employment Agency and contain information on an individual's employment (with the exception of self-employment) and unemployment episodes, on socio-demographic characteristics, on transfer payments (unemployment benefits) and on participation in active labour market policies. IEBs have a long history and it has required continuous effort and investment to develop and manage the dataset. More details on IEB are provided in the box 7[37].In the German evaluation of job creation schemes, these data were complemented by a survey structured in three

[36] These are "eligible but not admitted" participants; non-admission generally depends on the end of funding or some other external causes (e.g., sickness of teachers, transportation breakdown, etc.). These subjects can represent a preferential control group because they have the same eligibility and the same will to participate as the actual participants; however, there are sometimes not enough of them to generate a control group, or their details are not registered.

[37] A similar experience is represented by the Irish database "Jobseekers Longitudinal Dataset" (JLD), which integrated information on payments for social assistance and social insurance, labour market programmes, and employment histories in a single database. Among others, JLD was used in the evaluation of the JobBridge activation programme financed in the programming period 2007-2013. See Indecon, 2016.

waves to collect information on the "soft outcomes" (as perceived by the individuals) at different stages during the interventions.

In the Italian evaluations examined in detail in the Guide, the data used were similar. Archives of the unemployed registered at PESs were used to identify potential control groups, while administrative data on employment changes (Comunicazioni obbligatorie - COB), registering all the labour contracts of firms and public employers (excluding self-employment), were used to measure the outcome variables. As COBs are managed at regional level information quality varies. However, nationwide standards have been introduced in the last few years, and quality has improved. In one evaluation (for the Province of Trento), the evaluator was also able to merge these data with the tax return data provided by the INPS (the national institute of social security), so being able to measure the impact in terms of an individual's earnings.

In Poland evaluations used administrative data drawn from the PESs unemployment registers to identify control groups. However, these data do not contain information on the employment status or history of individuals, and evaluations had to rely on a proxy (cancelling unemployed status at the PESs), or to collect information through surveys of a sample of individuals, both treated and not treated.

In Latvia, the evaluator was able to merge two main administrative datasets. Data from the Latvian State Employment Agency (SEA) provided information on both participants and non-participants registered as unemployed on specific dates, and data from the State Revenue Service (SRS) provided information on employment conditions on different dates as well as on the income of individuals. This allowed the evaluator to assess the effects both in terms of likelihood of being employed at different times and also in terms of income.

The only case of randomisation, the Swedish example, used both administrative data to measure outcome variables and a survey of employment offices and intermediaries to measure the intensity and types of support provided.

The evaluations in the field of education used administrative data to identify the control group and to assess the outcome variables, with the exception of Poland, which had to rely on a survey to measure the variables used as outcomes.

Overall, the examples show the importance of appropriate administrative data for CIEs, ideally both for the identification of control groups and the measurement of outcome variables. From this point of view, MAs planning CIEs should ensure in advance that the administrative data for carrying out the evaluations is available, putting in place appropriate actions to tackle potential issues of accessibility, integration of data or other problematic aspects.

Box 7 Examples of integrated databases for CIEs

The Jobseekers longitudinal dataset (JLD) in Ireland

The Jobseekers Longitudinal Dataset (JLD) is an administrative database managed by the Department of Social Protection (DSP).

JLD is an ambitious attempt to adapt administrative data to research purposes. Development started around 10 years ago, after the DSP had commissioned University College, Dublin to carry out a preliminary study on the management of the Live Register[38] and, more generally, data relating to the labour market. The study provided several suggestions for improving data collection and identified some challenges (for example duplication of data in various data systems, missing information, …).

JLD integrates several sources of information: payment and administrative data from the DSP – e.g., payments for social assistance and social insurance to the working-age population included in the Live Register and data related to active labour policy programmes managed by DSP; data on labour market programmes managed by SOLAS, the national education and training body; data collected by the tax authorities (Revenue Commissioners). With regard to ESF interventions, JLD covers partly the interventions financed by the fund, but more specifically those financed by the DSP and SOLAS[39].

[38] The Live Register contains information on people registering for Jobseekers' Benefit (JB) or Jobseekers' Allowance (JA) or for various other statutory entitlements at local offices of the DSP.

[39] Programmes in areas such as Justice (the Youth Diversion Projects and Young Persons Probation Projects, as well as Integration and Employment of Migrants), Education (Third Level Access and Adult Literacy), Community (Community Training Centres), and

Data from the aforementioned sources are rearranged as a series of episodes, with one episode starting when the person begins a spell of unemployment and ending when the person moves into employment or another activity or training programme; then, when the status of an individual changes again, a new episode begins. The beginning of an unemployment period coincides with an individual starting to claim Jobseekers' Benefit and Jobseekers' Allowance[40]. In the JLD, an advantage of this structure by episodes is that contiguous periods on Jobseekers' Benefit (JB) and Jobseekers' Allowance (JA) can be linked and represented as one episode of unemployment. It is worth noting that episodes, as seen above, can overlap and the researcher has the problem of prioritising one episode over another.

JLD has been tracking social welfare claims, activation, training, and employment histories since 2004 and includes approximately 13 million individual episodes of welfare and work for around 2 million individuals. Each episode has a start and an end date, and an operational code[41] which permits the identification of the situation of an individual during each spell. JLD covers a large set of 'variables': gender, age, marital status, citizenship, educational attainment, previous occupation, employment and unemployment histories (duration and number of episodes) and characteristics of jobs (sector for example), unemployment training history (type, duration and number of episodes), benefit type (JA, JB), number of children, dependents, family payment type (i.e. adult and child dependent allowances, adult only, etc.), earnings and tax and geographic location[42].

Furthermore, individual identifiers link JLD to other administrative data; for example, for the evaluation of JobsPlus JLD was matched to a separate monitoring database, with detailed information on JobsPlus start/end dates, number of days on the Live Register at the beginning of JobsPlus, and type of treatment.

There are two main channels for accessing JLD data: contracted research, which occurs when DSP launch a Request for Tender (RFT) process; or on request from researchers with adequate credentials. In this latter case access requests are considered on a case-by-case basis, and in case of acceptance, a legally enforceable data-sharing agreement between the researchers/institutions and DSP must be signed. All data are pseudonymised, and only the data fields relevant to the research topic are transmitted. Access to data by the researcher or research institution is limited, and data must be deleted at the end of the research or evaluation project. A codebook is provided to researchers along with JLD data, though significant improvement in documentation is needed.

JLD has been used for a variety of research activities and evaluations over the years: Back to Education Allowance (2015), JobBridge Activation Programme (2016), Back to Work Enterprise Allowance (2017), JobPath (2019) and JobPlus (2020) are some of those which have been selected for evaluation.

JLD structure and content have improved since the initial phases. However, it requires constant activity to keep it updated and to develop improvements and fill the gaps, for example by including more detailed information on education and integrating more programme-specific data (content, completion etc.), and improving earnings data. An upcoming project should implement improvements to JLD in terms of: regular updating (for example monthly) with automation and testing of the data pipeline, and adding or replacing data sources to ensure that there is full coverage of employment and training support, other social protection programmes, and real-time earnings data[43].

The Institute for Employment research (IAB) in Germany and the Integrated Employment Biographies (IEB)

In Germany, data access for scientific purposes was improved following the 2003-2005 labour market reforms, putting emphasis on the evaluation of policies and the recommendations of the Commission on 'Improving the Information Infrastructure Between Science and Statistics' to establish a research data centre

Defence (Defence Forces Employment Support Scheme) are not included in JLD.

[40] Jobseekers' Benefit is a weekly payment from the DSP to people who are out of work, who are fully unemployed or are in part-time employment because their work hours were reduced by their employer. It is applicable to jobseekers who have paid social insurance (PRSI) at the appropriate rate who have sufficient contributions in the relevant tax year, and have paid a minimum total of 104 contributions. Jobseekers' Allowance is a means-tested payment made to jobseekers who are unemployed and do not qualify for Jobseeker's Benefit or whose entitlement to Jobseekers' Benefit has expired.

[41] Representing a combination of activation/training activities, welfare claims, and time spent in employment.

[42] An element which requires attention is that some data refer to different periods, for example earnings are registered on an annual basis, while the social protection benefit payments are updated weekly.

[43] We would like to thank Frank Humphreys, Ciaran Judge, Saidhbhín Hardiman and Krzysztof Gigon of DSP for the information provided on JLD.

at each public producer of microdata. As a result, the Federal Employment Agency established a research data centre within the Institute for Employment Research (Institut für Arbeitsmarkt- und Berufsforschung (IAB)) in 2004. IAB is responsible for extracting data from administrative processes to prepare datasets useful for empirical research.

More than 15 datasets are managed by the IAB and are available to the scientific community. The data originate from administrative data from the social security system, internal processes of the Federal Employment Agency and from surveys conducted by the IAB[44].

In relation to the social security systems, all employers are required to report a number of items and characteristics of their employees and these data provide a rich set of information about people's employment history. The administrative data relating to the internal procedures of the Federal Employment Agency include compulsory unemployment insurance, unemployment benefits and the corresponding entitlement periods, advisory meetings with unemployed individuals, placement offers, and active labour market measures. The IAB combines the data into a single comprehensive dataset, which is called the "Integrated Employment Biographies (IEB)". The collection of these administrative data began in 1975, though not all variables are available for the entire observation period. The IEB can be integrated with several data from surveys implemented by IAB, such as: IAB Establishment Panel, the German Job Vacancy Survey of the IAB, the German Management and Organizational Practices (GMOP); Panel Study Labour Market and Social Security (PASS); Working and Learning in a Changing World; Employee survey bonus payments, wage increases, and fairness (BLoG); Linked Employer–Employee Data from the IAB (LIAB); Panel "WeLL"—Employee Survey for the Project "Further Training as a Part of Lifelong Learning".

IAB updates its data products regularly and offers different samples of these rich administrative data sources for research purposes. For each data product, IAB provides detailed documentation in German and English. The legal basis for data access is found in the German Social Code Book (several versions) and, more specifically, four kinds of data access for the scientific community are envisaged: Campus files (fully anonymised and useful only for teaching); Scientific Use Files are anonymised microdata submitted to scientific institutions in Germany and EU Member States for research projects in the field of labour market research but not for teaching or for commercial research interests. Data security must be guaranteed by the scientific institution applying for the data; weakly anonymised data with more detailed information are accessible only on-site. IAB provides separate workplaces within a secure computing environment in Nuremberg and at various locations in Germany, the USA, and the UK. Researchers have direct access to data, but they can obtain the output of their programmes only after disclosure reviews by IAB staff; remote execution means that researchers prepare their programmes with artificial data and upload them in the Job Submission Application (JoSuA). In this process, researchers never view the original data and they receive only their results. Standardised request forms are available for all data access. After a request has been approved, a contract governing data use for a specific project within a specific period is concluded between the researcher's institution and the IAB. The contract specifies the data protection rules, and severe sanctions apply in the event of violation. Some of the datasets managed by IAB are available only for on-site use (for example the linked datasets).

Based on data produced by IAB, active labour market policies are evaluated regularly using the latest empirical methods and in some cases the findings have led to a change in the policies[45]. IAB provides researchers with access to its datasets not only in Germany but also in other countries and the number of users is steadily increasing (for example, in 2016 almost one-third of all data use agreements were from a non-German facility).

On the use of administrative data for the CIEs and related practical issues see also the European Guide prepared by researchers from JRC, European Commission, 2020.

[44] In 2011, the Record Linkage Centre was founded at the IAB, a joint project with the University of Duisburg-Essen that was funded by the German Research Foundation; the Centre aims to simplify the linkage of datasets without a specific identifier.

[45] An example is an evaluation of the compulsory integration agreement between the jobseeker and the caseworker. Using a randomised field experiment and following the labour market biographies of the persons included in the experiment, IAB was able to show that for some groups of unemployed, the compulsory regulation is counterproductive and should be replaced by a more flexible handling of the instrument (van den Berg et al. 2016).

What are the possible data protection issues?

Difficulties can be experienced in obtaining data that identify individuals or companies who have participated in ESF+-financed interventions[46].CIEs require micro-data - data which contain observations on the individual units in both treatment and control groups.

ESF+ Regulation on data management

ESF+ Regulation 2021/1057 (Annex I)) asks for data on participants with a breakdown by gender, labour market status, age group, educational attainment, and vulnerable groups (migrants, minorities, disabled, other disadvantaged). The CPR and ESF+ Regulations for 2021 - 2027 establish a legal obligation for MAs to collect and process personal data in the form of individual participant records. In addition, ESF+ Regulation 2021/1057 at Article 17(6) in relation to Monitoring and Indicators specifies that: "*Where data are available in registers or equivalent sources, Member States may enable the managing authorities and other bodies entrusted with data collection necessary for the monitoring and the evaluation of general support from the ESF+ strand under shared management to obtain data from those registers or equivalent sources, in accordance with Article 6(1), points (c) and (e), of (EU) 2016/679"*.

These rules, set out in the Common Provisions and ESF+ Regulations, facilitate access to and use of personal data needed for ESF+ monitoring and, in the case of a CIE, for defining the treated group. However, access to personal data necessary to form the control groups and the processing of treatment and control group data must be in line with Regulation (EU) 2016/679 (General Data Protection Regulation – GDPR), which covers the general transfer and use of personal data, including special categories of data[47] within the EU. The following box provides an overview of the main contents and obligations regulated by the GDPR.

Box 8 EU regulatory Framework on personal data processing

Relevant legislation on the processing of personal data in Europe consists mainly of Regulation (EU) 2016/679 and the guidelines and measures adopted by competent authorities such as the European Data Protection Board (EDPB). This legislation sets out many conditions and limitations on the processing of personal data to protect the rights and freedoms of data subjects. At all events, the need for protection and safeguard that emerges from the provisions and regulations has to find a balance with the need not to constrain scientific research and indeed to act as an asset in its development. For this reason, and within this legal framework, the European Union has foreseen that, under specific conditions, exceptions can be made to allow research activities and the dissemination of the outputs as long as the first and foremost right of the persons concerned, i.e., the right to privacy, is safeguarded.[48]

GENERAL DATA PROTECTION REGULATION (EU) 2016/679

The General Data Protection Regulation (GDPR) came into force on 24 May 2016 and became fully applicable in all Member States on 25 May 2018.The GDPR applies "*to the processing of personal data wholly or partly by automated means and to the processing other than by automated means of personal*

[46] See Ismeri Europa – Ecorys – Institute for Employment Studies, 2019.

[47] The GDPR no longer uses the term sensitive data and now refers to 'special categories of data'. These, according to Article 9, include: "data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation".

[48] Other relevant European legislation that contributes to the legal framework for the processing of personal data includes: Regulation (EU) /2013/557 on European Statistics as regards access to confidential data for scientific purposes; Regulation (EU) 2018/1725 on the protection of natural persons with regard to the processing of personal data by EU institutions, bodies, offices and agencies and on the free movement of such data.

data which form part of a filing system or are intended to form part of a filing system"[49].

As regards territorial scope, the Regulation applies both to the processing of personal data carried out by persons located in the territory of the European Union and in cases where processing involves data subjects located within the European Union, even when the data controller or processor is located outside the EU.

From a substantive point of view, and compared to the previous legislation, the Regulation reinforces the rights of data subjects and imposes a series of obligations on data controllers according to a logic based on risk analysis and on the principle of accountability. Moreover, the Regulation provides for a series of requirements to be fulfilled by the data controller, requirements which were not included in the previous legislation: among these, the privacy impact assessment (Articles 35-36), the minimisation of processing operations according to the criteria of privacy by design and by default (Article 25), the adoption of the processing register (Article 30) and the appointment of the Data Protection Officer (DPO) (Articles 37-39).

It is important to highlight the main principles set out in Article 5 of the Regulation, with which those who process personal data must comply. In particular the following:

a) *Lawfulness, fairness and transparency*: personal data shall be processed lawfully, fairly and in a transparent manner.
b) *Purpose limitation*: personal data shall be collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes. In the case of processing for statistical purposes or scientific research, the data controller shall adopt the necessary guarantee and protection measures.
c) *Data minimization*: only data strictly necessary to achieve specific purposes should be collected.
d) *Accuracy*: the data collected shall be accurate and, where necessary, kept up to date.
e) *Storage limitation*: personal data shall be kept in a form that permits identification of data subjects for no longer than is needed for the purposes for which the personal data are processed. To this aim, personal data may be stored for longer periods insofar as the data is processed solely for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes in accordance with Article 89. In such cases, the data shall be subject to the implementation of appropriate technical and organisational measures to safeguard the rights and freedoms of the data subjects.
f) *Integrity and confidentiality*: personal data shall be processed in a manner that ensures appropriate security thereof, including protection against unauthorised or unlawful processing and accidental loss, destruction or damage, using appropriate technical or organisational measures.
g) *Accountability*: is one of the most important principles set out by the Regulation. It states that the controller is responsible for data processing and can demonstrate the implementation of any required measures.

This set of principles constitutes the main structure on which the GDPR is based and determines a series of obligations to be fulfilled by the controller and the processor.

PROCESSING OF PERSONAL DATA IN THE CONTEXT OF THE CIE: CONDITIONS, LIMITS AND MAIN ISSUES

Conducting a counterfactual impact evaluation involves processing large amounts of data, including personal data. Referring to the regulatory framework described above, it is important to be familiar with the conditions, limitations and main issues that MSs and MAs face when conducting CIEs.

Legal basis and purposes of the processing

To carry out the processing of personal data, at least one of the conditions of lawfulness indicated in Article 6 of the GDPR must be present[50]. If, on the other hand, the data to be processed are "special categories of data", reference must also be made to Article 9 GDPR.

In the context of the CIE, **public interest** according to art.6(1) e) of the GDPR ("*the performance of a task*

[49] Art. 2, Regulation (EU) 2016/679 of the European Parliament and of the Council.
[50] These conditions are: a) the data subject has given consent to the processing of his or her personal data for one or more specific purposes; (b) processing is necessary for the fulfilment of a contract to which the data subject is party or in order to take steps at the request of the data subject prior to entering into such contract; (c) processing is necessary for compliance with a legal obligation to which the controller is subject; (d) processing is necessary in order to protect the vital interests of the data subject or of another natural person; (e) processing is necessary for the performance of a task carried out in the public interest or in the exercise of official authority vested in the controller; (f) processing is necessary for the purposes of the legitimate interests pursued by the controller or by a third party, except where such interests are overridden by the interests or fundamental rights and freedoms of the data subject which require protection of personal data, in particular where the data subject is a child.

carried out in the public interest or in the exercise of official authority vested in the controller") appears the most appropriate legal basis for data processing. This legal basis has to be laid down in EU or national law, as specified in art. 6(3) of the GDPR. The 'public interest' clearly represents the obligations of the managing authorities defined in ESF+ Regulation 2021/1057 art.17(6) on the use of data "*available in registers and equivalent sources*". In addition, CPR 2021/1060 states in art. 4 that "*the Member States and the Commission shall be allowed to process personal data only where necessary for the purpose of carrying out their respective obligations under this Regulation, in particular for monitoring, reporting, communication, publication, evaluation […]"*. National laws can also vest data controllers with similar authorities for managing and processing data in the public interest.

Other legal bases for a CIE may be referred to in other conditions listed in art.6(1) and in particular:

- the **consent** of the data subject (art.6 (1) a) GDPR). The consent, for instance, may be an appropriate legal basis when data for the CIE are collected by a survey and the data subjects can easily give their consent to the processing (see art.7 of GDPR on consent). In general, consent is more complex if not planned well in advance; it may be considered a "residual" legal basis for CIE when other legal bases are not applicable.
- processing is necessary for **compliance with a legal obligation** to which the controller is subject (art.6 (1) c) GDPR in reference to art.17 of ESF+ Regulation). This legal basis has to be established by law and may involve private or public entities; it could be, for instance, that due to specific legal prescriptions, a private or public entity responsible for a dataset is required to collaborate with the managing authority in the CIE.

In addition, the data controller **may use the collected data for further purposes if these are compatible with the initial purposes**. In this regard, the data controller will have to evaluate the conditions laid down in Articles 6(4) and 5(1) b) of the GDPR. These provisions are particularly relevant when processing data for scientific research or statistical purposes. They allow the use of administrative data for purposes that differ from the original ones and do not require specific consent to the new use, but are required to comply with the protection rules specified in art. 89(1), mainly pseudo-anonymisation (see below). These provisions, for instance, may be relevant in the case of unemployment register data to be used in a CIE.

Due to the specificity of CIEs, the data controller collecting the data and the entity which carries out the research may be different. In this case, there must be a condition that legitimises the data transfer and allows the recipient to proceed with the CIE.

Anonymisation and pseudonymisation: processing for statistical purposes

Article 89 of the GDPR states that data processing carried out for purposes of public interest, or in the context of scientific research or for statistical purposes shall provide for appropriate safeguards for the rights and freedoms of data subjects and shall respect in particular the principle of "minimisation". This means making use of pseudonymisation techniques[51].

Where the purposes can be fulfilled by subsequent processing operations which do not permit, or no longer permit, the identification of data subjects, those purposes shall be fulfilled by anonymising the data and then processing them in aggregated form. In CIEs, by definition, results are aggregated and this risk is absent, unless the original datasets are published for scientific reasons. In this case, datasets must be anonymised (see the example in the box below).

Data storage and secure processing

One of the main aspects of data processing is data storage. The legislation does not specify how data must be stored, but the principles mentioned above require that storage and processing always be linked to the purpose of the research. When the purpose of the processing is achieved and retaining the data is no longer necessary, it shall cease. This general rule must be specified in the privacy statement that is given to data subjects when they are registered in administrative datasets or, given the likely impossibility of informing all data subjects personally, alternative ways of providing the information may be found (for instance, by publishing an information page containing the privacy policy on the research activities on the website of the

[51] Pseudonymisation means the processing of personal data in such a manner that the personal data can no longer be attributed to a specific person without the use of additional information. Anonymisation refers to the processing of personal data in a manner that makes it impossible to identify individuals from them. An overview of pseudonymisation techniques can be found at https://www.enisa.europa.eu/publications/pseudonymisation-techniques-and-best-practices.

managing authority).

In particular, once the CIE has been carried out, the results made available and aggregated for statistical use, the purpose is considered fulfilled. At this point, data should either be anonymised (in case they are to be used again at a later stage and for different purposes) or deleted. Law or regulation permitting, storage may be unlimited in time, but only under explicit regulatory reference.

The other condition required by the GDPR is that processing be surrounded by adequate and appropriate security measures. Article 32 of the Regulation states that whoever carries out the processing (controller or processor) shall implement appropriate technical and organisational measures to ensure a level of security appropriate to the risk. This means that no standard measures are prescribed, but that they must be defined on a case-by-case basis with specific reference to the risks posed by the individual processing operation.

LEGAL OBLIGATIONS

Information for data subjects

The first duty of the data controller is to inform the data subjects. This obligation is set out in Article 13 GDPR when the data is collected from the data subject, but when this is not the case the reference is Article 14. The subject must be informed about the purposes and methods of the processing, the legal basis, the storage periods and the rights that can be exercised by the data subject. In the case of CIE, this information is given at the time of collecting the data which will be later used for statistical and evaluation research, such as in ESF monitoring or unemployment register datasets. When this communication is practically impossible, as mentioned above and especially in the case of individuals in registers used for the CIE, website or other general tools of information can be used.

Governance of the relationship between the different entities involved in processing

Generally, research activities involve more than one body and, in such cases, the relationships between them must be regulated by specific arrangements known as “data processing agreements”[52]. These are to be defined on a case-by-case basis reflecting the contributions of the different stakeholders. Examples of possible relationships are:

- *Data controller – Data controller* occurs when the entities collaborate in the implementation of a project, although in different conditions and with different tasks, each maintaining its own distinct processing purpose;
- *Data controller - Data processor* occurs when one entity (Data controller) determines the means and purposes of processing and uses another entity (data processor) to perform certain processing activities. In this case, the reference and obligations are to be found in Article 28 GDPR;
- *Joint controllers*: this type of relationship is set out in Article 26 of GDPR, which provides that: “*Where two or more controllers jointly determine the purposes and means of processing, they shall be joint controllers*”. In such cases, the parties must determine their respective responsibilities for compliance with the regulations in a transparent manner, in particular with regard to the rights of the data subjects.

Data protection impact assessment

Article 35 of the GDPR states that, where the processing of personal data is likely to present a potentially high risk to the rights and freedoms of data subjects, the controller must carry out an impact assessment[53] before proceeding with the processing. An impact assessment in CIEs may be important only in case of processing special categories of data subjects on a large scale. In such cases, the data controller in the administration may conduct the impact assessment according to the rules and the tools made available by the national Data Protection Authority.

FUTURE PROSPECTS: THE DATA GOVERNANCE ACT

The framework outlined above reflects the current state of the art and the rules in force in the EU. The GDPR was introduced specifically to enable a breakthrough in personal data protection, and to create a common system across the EU in accordance with current technological developments and today’s data society. Other reforms are in the pipeline and may have a significant impact on the collection of personal data in the coming years.

[52] For instance, this is the case when the data controller and those actually conducting the research are different; In this case it is necessary to find an agreement that regulates relations between the parties to access data and carry out the research.
[53] For further information, consult the guidelines published by the EDPB https://ec.europa.eu/newsroom/article29/items/611236.

The Data Governance Act is currently still at the proposal stage in the European Commission[54]. The explicit aim of this act is to promote the availability of usable data by strengthening trust in data intermediaries and enhancing data sharing mechanisms across the EU. Personal data are likely to be the subject of the new standard, and their use for statistical and research purposes could be more far-reaching thanks to the introduction of a new figure: the personal data sharing intermediary. Such a measure could contribute to creating a more enabling environment for CIEs, allowing easier access to a large amount of information on a large scale.

Main steps to better comply with GDPR

It is clear that the rules in the GDPR do not rule out the implementation of a CIE; as a result, some basic procedural and operational steps must be respected in all MSs in order to make the CIE secure. For example, in a normal case where monitoring data (treated people) are combined with data from a public register (control group) and an external evaluator carries out the CIE, the most important steps are:

1. The MA comes to an agreement with administrations responsible for data (e.g., unemployment register, tax register, etc.) necessary to identify the control group and analyse treatment and control groups. The MA verifies that the use of this data complies with art. 6(1) e), or art. 6(4) and 5(1)b) of GDPR, if expressed consent was not collected;
2. The MA makes agreements with the other entities (data owners and the evaluator) to regulate the flows of information and the mutual responsibilities in accordance with GDPR rules. The service contract between the evaluator and the MA must include a specific clause on data protection; in the case of other administrations, a memorandum of understanding or specific national procedures can regulate data protection when implementing CIEs.
3. On the basis of the above agreement, the MA receives data in pseudo-anonymised form from the data-owner and transfers data for processing treated and control groups to the evaluator in compliance with the GDPR[55].
4. The storage of personal data of treatment and control groups only complies with the rules of data storage for the duration and purposes of the research and in accordance with basic security rules, as stated in the GDPR. These rules are observed by all the entities involved in the CIE.

National practices are generally consistent with GDPR, but differ among MSs, and evaluators report that national data protection rules still pose serious obstacles to the use of micro-data. This may stem from the time required to adapt national rules and habits to the more recent GDPR, or from different interpretations of GDPR in different national administrations, gold-plating, making the GDPR stricter in some countries, or other misinterpretations. In some countries a specific initiative of the MAs or the national authorities would be necessary to overcome these obstacles in the spirit of the GDPR and ESF+ Regulation. How the Veneto Region in Italy dealt with some of the most common difficulties is explained in the Box below. A good practice is to get in touch with the national Data Protection Authority to discuss the proposed arrangements before finalising them.

[54] The proposal for a Data Governance Act was adopted at the end of 2020 by the Commission. The state of play of the Data Governance Act can be found here: https://eur-lex.europa.eu/legal-content/EN/HIS/?uri=CELEX:52020PC0767.
[55] If treatment and control groups are analysed by survey, this has to include consent to use the data of the interviewed people for research purposes.

Box 9 Data protection and exchange

A good example of accessing anonymised personal data in a relatively short time can be found in a counterfactual evaluation in the "Pilot and feasibility study on the sustainability and effectiveness of results for European Social Fund participants using counterfactual impact evaluations"[56] carried out for DG EMPL in 2019.

The evaluation, one of four cases included in the study, focused on ESF interventions (IP 9.i) carried out in the Veneto region in 2015/2016 targeting the long-term unemployed (LTU). The monitoring data on participants and the types of measures were combined with data on employment spells before and after the ESF interventions (called "Comunicazioni obbligatorie"), The process lasted about two months (from the end of July to the end of September) from the first meeting on data requirements with the managing authority and the institution holding the data, the regional Institute of Veneto Lavoro. Veneto Lavoro was responsible for the anonymisation of the data and all the datasets provided were easily linkable to one another[57] by means of a unique common identifier provided by Veneto Lavoro.

Though the organisations involved were not directly commissioning the evaluation, the positive experience is due to two main factors: strong (albeit informal) cooperation between the regional offices of the managing authority and the regional Institute of Veneto Lavoro, necessary to address and solve the data protection issues, and the existence of a database (Mercurio) which is managed by Veneto Lavoro and is available as a public use file for research purposes[58].

Mercurio is a statistical database containing all details registered by labour exchange offices in the Veneto Region regarding spells of employment and unemployment in the population. Moreover, it contains detailed information on all registered workers and firms. At the time of the analysis Mercurio contained information on more than 4 million workers, 17.6 million employment events and 4 million unemployment events[59].

The dataset is periodically "updated and cleaned" and this made matching it with the monitoring data of the MA relatively straightforward. One potential limitation is the fact that the public use file is not updated often enough (for example the last version currently available is updated at December 2020), since it requires huge effort to ensure good quality of data[60]. Experience suggests that the MA should make arrangements in advance, while preparing evaluation plans for: potential forms of cooperation with external bodies managing the necessary data for a counterfactual exercise; solutions to potential legal obstacles; clarifying ways to access, manage and provide administrative data in anonymised form.

2.3.2. How is the 'treated' group to be identified?

In order to conduct a CIE, a clear definition of what it means to be treated or to have participated in the intervention is essential. Moreover, once a clear understanding has been reached as to when an individual or enterprise is said to have been treated, it is important that they can be identified. Here the main issues relating to the definition of treatment and control groups are introduced, Chapter 3 provides more detailed explanations of their methodological characteristics.

Definition of the treatment group

Defining participation might appear straightforward at first sight. However, there are a number of issues that may not be immediately apparent but which are crucial and require careful thought. For example, are trainees in a training scheme that drop out of the intervention considered to have been treated? How many sessions in a training course do trainees need to attend before

[56] Ismeri Europa – Ecorys – Institute for Employment Studies, 2019.
[57] An additional dataset containing the beginning and end dates of each Declaration of Immediate Availability to Work signed by all treated and not treated individuals was provided.
[58] https://www.venetolavoro.it/public-use-file for a description of Mercurio and the rules governing its access.
[59] The following variables of interest were extracted from the administrative data: Dates of the beginning and end of each employment and unemployment spell; type of contract for each employment spell (permanent, temporary, internship); expected duration of temporary contracts; socio-demographic information (gender, date of birth, education, citizenship)
[60] This means that it is necessary to work on more updated information than that contained in the latest available version of Mercurio, specific extra work is needed to integrate and clean the most updated information on employment spells.

they are considered participants? There are also anticipatory effects to consider. In anticipation of being subject to an intervention, some claimants of social security benefits may leave welfare rolls in order to avoid activation measures. Are these individuals treated even though, for example, they never physically attend appointments made for them at a PES office?

Intention to treat or treatment of the treated

There is a distinction between 'intention to treat' and 'treatment of the treated' when defining the 'treatment group'. From a policy perspective, the key question is usually whether interest lies in the effects of being offered the opportunity to participate in an intervention, or the effects of actually participating. In the former case, those offered an intervention may or may not participate. In the latter case, where interest is in the effect of treatment on the treated, the treated group contains only those who participate.[61]

Offer or actual treatment

At first glance, policy-makers often assume that they are interested in determining the net effects of treatment on those who participate. However, on further reflection, the issues can be less clear cut. If those who are offered treatment can be identified, it may be more useful from a policy perspective to define them as the 'treated' group. This is particularly so in circumstances where participation in an intervention is non-mandatory. Policy-makers cannot force those offered an intervention to participate, so the relevant question is: what is the impact on subsequent employment and earnings for those who were offered the opportunity to take part in a training programme?

To estimate the effects of the offer of treatment on a range of results, those who receive the offer need to be identifiable. In many cases this might be difficult.

Where to find suitable data

Finding data sources for treated persons

Once definitions of who are treated and what constitutes treatment have been decided, it is important to consider how those who are treated will be identified for the purposes of the evaluation. This invariably means finding a data source from which treated units, be they persons or enterprises, can either be fully enumerated or sampled. These records are usually drawn from the ESF monitoring systems and - if available – from further data records established for the particular intervention.

Due to ESF (and ESF+) monitoring and reporting requirements, beneficiary organisations need to record the numbers and some personal characteristics of those who receive services through an intervention. For the purposes of CIE, interventions will need to go further and provide micro-data on those who have participated. Evaluators will require a record for each treated unit (enterprise or person) with data on their main characteristics (sex, age, level of education, etc.). These data can be anonymised/pseudonymised for privacy purposes, but when a survey is needed to carry out the CIE, it will be necessary to establish whether conditions required by GDPR allow the transmission of the identities of these units (names, addresses, telephone numbers, etc.) so that they can be sampled. Unique identifiers for each individual unit are also required to facilitate the linking of records across data sources.

[61] Where participation in an intervention is mandatory, there is essentially no difference between these two statuses-everyone offered treatment has to participate. However, in most cases interventions are non- mandatory (as assumed throughout this Guide).

2.3.3. Factors to be considered in identifying a control group

Three aspects to take into consideration in defining a control group

To obtain an estimate of the counterfactual, a control group will usually need to be identified. The choice of a control group will usually be constrained by whether the intervention is mandatory for participants or not, as well as whether the intervention is implemented universally within a jurisdiction, or limited to a particular area or over a limited time span. The choice of an appropriate control group has three aspects: 1) analytical; 2) policy-related; and 3) practical.

Defining a control group from an analytical perspective

The purpose of CIE is to obtain unbiased estimates of the impacts of an intervention on a range of results. To achieve this, estimates of counterfactual results are required. Counterfactual result estimates are obtained from a control group (see Section 1.1). As shown in Figures 6 and 7, an impact is estimated by subtracting an estimate of the counterfactual result from an observed result for the 'treated' group. The extent to which an impact is biased depends on the degree to which the counterfactual result computed for the control group represents the result which would have emerged for the treated group had they not been treated, all else remaining equal.

To find an equivalent control group in the absence of randomisation, it needs to be equivalent to the treatment group in all important respects, both in observable and unobservable dimensions. This is true in all the quasi-experimental approaches and, consequently, a necessary condition for all the CIEs not adopting a randomised approach.

Options for selection of control groups

As almost all ESF interventions are either a) voluntary (the target group are not compelled to participate in an intervention), and/or b) limited in some other way – If they are pilot interventions or instruments restricted to a particular region or jurisdiction, evaluators will be confronted with a pool of units that could be selected for use as controls. Some sifting of this potential pool will be required in order to refine the final choice of controls so that they are well matched to participants (the treated group). In many cases, four options are potentially available:[62]

- **Location -** controls that are similar to those participating in an intervention but located in areas of the MS where the intervention is unavailable (should such areas exist). Difference-in-differences is often the favoured approach in the case where such control groups and the right data are available. Populations in different locations can be very similar to each other and such groups will not have had the chance to participate in the intervention or decline the invitation to do so, and therefore this important source of potential bias will be absent. However, populations in different locations will be subject to different labour market conditions. Difference-in-differences techniques for such variations work quite well, as differences in local labour market conditions tend to be reasonably fixed over time. It is less advisable, however, to draw control samples from different local labour markets in cases where matching is being used to estimate impacts. It has been shown that the bias associated with selecting control samples from different labour markets can be greater than selection bias;[63]

[62] This section draws on Card, D., Ibarraran, P. and Villa, J.M., 2011.
[63] See Heckman, J.J., Ichimura, H., Smith, J. and Todd, P., 1998.

- **Time -** controls that are similar to participants but are observed at different points in time, either before or after the intervention. Control groups selected in this way are often required where an intervention is universal and mandatory – in other words, where all target group members are obliged to take part and the programme is implemented across an entire jurisdiction. Control groups formed in such a way possess a significant disadvantage, namely that their results will be measured at different time points to those of the treatment group, thus being susceptible to cyclical fluctuations, compositional changes and shifting macroeconomic trends that may stifle the capacity to identify an unbiased counterfactual result. Such controls should only be considered where there is limited variation in results over time, and where a contemporaneous control group is unavailable;

- **Eligibility –** here, controls are selected from groups at the same location and point in time but from amongst candidates who for one reason or another were ineligible to participate. Such controls are often sought where an intervention is universal, participation rates are high, or participation is mandatory and where there are clear eligibility rules, such that, for example, those 'just ineligible' provide a potential source of controls. The objective is to find groups who are similar to those treated but who for well-known and fixed reasons (which can be quantified in the data) were not eligible for treatment. Access to interventions under ESF+ is often based on distinct eligibility rules that can be readily measured and are not open to manipulation (specific age of the participants, duration of unemployment status, etc.); therefore, the selection of controls can use these thresholds as discriminating factors around which treated and controls are distributed.

- **Choice/awareness -** In essence, both treatment and control groups (rather than just the treatment group) are subject to selection processes based on choices motivated by potentially unobserved factors.[64] Controls can be selected from amongst those who were eligible but failed to participate. The advantage of this is that they are usually drawn from the same labour market as those who were treated. Such controls should be considered carefully, especially where a matching CIE design is being used and where rich data can be drawn upon to inform the selection decision. In other circumstances, for example where difference-in-differences is being implemented, choice/awareness controls will be less attractive.

Analysing pre-intervention trends

One further point is noteworthy. Where pre-treatment result measures are available for both the treatment and control groups, it is important to examine pre-intervention trends in result measures for both groups. Checking the so-called 'common trends' assumption addresses the problem of temporary pre-intervention dips in employment rates and earnings that will have occurred for some of those eligible for ALMPs (otherwise they would not be eligible for support - the so-called 'Ashenfelter's Dip' problem). The evaluator is looking for similar time trends in result measures for both treatment and control groups, so that recovery from short-term job or earning loss will not be confused with the long-term relative gains CIE attempts to detect.

[64] This is what Card, D., Ibarraran, P. and Villa, J.M., 2011 refer to as 'two-sided selection' bias'.

The selection of appropriate control groups is a technically and methodologically complex exercise. During the development of evaluation schemes, it is recommended that officials make themselves familiar with the main concepts and take early steps to identify potential controls. It is important that commissioners of an evaluation engage experts early in the design stage to provide support and advice.

What are the relevant policy-related considerations?

Defining an alternative to an intervention

The selection of an appropriate control group is not simply a technical or analytical process. Though analytical aspects of identifying appropriate controls are fundamental, it is also important that a control group represent a relevant alternative to the intervention when considered from the perspective of policymaking.

Comparing treatment to no treatment or to an alternative

CIEs can take a number of forms: for example, they can compare the results of a treatment group or a number of treatment groups to a control group receiving no treatment; or they can compare one treatment to another without a no-treatment control group. The choice of control group will be informed by which type of comparison is most policy-relevant, and whether it is even possible to find a 'no treatment' control group. Box 10 below provides an example of a comparison of one treatment to another without a 'no treatment' control group – the objective being to assess whether to continue with one intervention rather than the other. It should also be noted that comparisons of one programme with another, without the benefit of a no treatment control group, can give rise to ambiguity (this is discussed further in Box 11).

Box 10 Policy questions related to a training programme

Consider an example where the policymaker intends to introduce a new training intervention which is to be funded through the ESF - call this Intervention A. Further, suppose that the MS already has a training scheme (B) targeted at the same persons but financed through national funds. In such a case a policy question might be: are the levels of employment and earnings for participants in intervention A greater than those for participants in Intervention B subsequent to participation? And by extension, does Intervention A represent better value for money? If earnings are higher for participants in Intervention A, then the obvious policy response is to discontinue Intervention B in favour of Intervention A, if its delivery also proves cost-effective.

Box 11 Interpreting net effects

A study may find no difference in earnings between participants in Intervention A and those in Intervention B. The policy response to this information may not be clear if, for example, Intervention B was highly effective relative to those receiving no treatment. This would mean that both interventions are highly effective. However, in some cases it might be that there is no evidence of the effectiveness of Intervention B relative to no treatment. Alternatively, interventions A and B could both be ineffective, though one intervention may appear relatively more effective than the other. In circumstances where certain groups in the population might be targeted by more than one intervention, it might still be more informative to find an appropriate group of untreated units to act as a comparison.

Note that difference-in-differences cannot be used to compare multiple treatments in the absence of a no-treatment control group.

What practical considerations are required for selection of a control group?

Alongside analytical and policy considerations, the practical aspects of selecting control groups need to be taken into account. Selecting or sampling units (persons or enterprises) to act as controls requires finding a suitable sampling frame. Furthermore, sampling frames should contain individual units that conform to analytical and policy requirements. Precisely how this is best done will vary from one evaluation to another, depending on the specific context of the intervention being tested.

In many cases, two sources of data are often exploited in identifying suitable control groups. Both require that the identity of the treatment group be known.

Population registers and company tax records

Population registers of various kinds can be used to find controls. For example, if an active labour market intervention is targeted at 18 – 24-year-old persons on unemployment benefit, their benefit records can be used to identify the target population. Further, if the treated group are known and can be matched to the benefits data, those 18-24year-olds who are untreated, and therefore potential controls, can be found. Alternatively, suppose an intervention is targeted at small and medium sized enterprises. National company records (where available) could be used to define the target population, and with information available on which enterprises are being treated, potential control groups found.

Applicant records

Applicant records can be used where take-up of the intervention is not universal; for example, where not all those who apply to a training programme are accepted (a choice/awareness control group). Similarly, not all those enterprises that apply for financing will be successful, and those not accepted for training or finance can in some cases be used as controls (see previous discussion in this section regarding the caution that should be exercised in selecting control groups under these circumstances).

2.3.4. What kinds of data issues need to be raised in the evaluation scheme?

What types of data are required and how will they be collected?

Managing data sources

As has been observed, CIEs usually require access to considerable quantities of micro-data (in some cases grouped data might be used – e.g., regional data). These data need to be collected, collated and documented; data from various sources need to be linked together on the basis of shared identifying fields; they need to be stored and transferred securely between those managing and undertaking the CIE; and analytical data sets need to be constructed from these data sources in order to facilitate estimation of impacts.

A basic check-list for data management

In developing an evaluation scheme, it is important to consider the following data-related questions:

Availability

– What sources can be used to obtain various types of grouped/micro- data?

- Are the necessary individual data available? Does this also apply to "special categories" data (if necessary)?
- When a survey is necessary, are target and control groups identified in a way that makes it possible to follow them up through survey interviews – are contact details available and up to date?

Consistency

- Is there one single data source or is it necessary to link data sources (e.g., statistics on unemployment, social benefits, social security, company/establishment data, etc.)?
- Are sources consistent with one another? Can individuals be identified within them on a consistent basis across sources?

Accessibility

- Is it possible to access national data sources on individual careers, earnings or social benefits to compare ESF participants with a potential control group?

Agreements

- Are specific agreements on data availability between the MA and other data owners in force? What legal or organisational barriers need to be discussed?
- Who will be responsible for negotiating access and obtaining agreement for data use?

Storage

- Where will data be stored? What IT systems and infrastructure will be required?
- What steps will be taken to ensure data are stored securely and that access is reserved for those who require the data for the purposes of evaluation?
- How are data anonymised? Is it possible to follow individuals through time and linked data sources?

Compliance with data protection

- Do previous solutions comply with the basic rules of the GDPR and the ESF+ Regulation? Are contingent obstacles in data accessibility justified by the use of different legal bases pursuant to art. 6 a), c) and e) of GDPR and Art.17 of ESF+ Regulation and the possible use of anonymised data?

How will data be processed?

Linking micro-data across sources

In a lot of cases CIEs will require micro-data – i.e., data which contains observations on individual units (usually individual persons or enterprises) in both treatment and control groups (occasionally grouped data might be used, e.g., regional or PES office-level data). We have distinguished between three main types of data required: a) treatment and control group records, b) result records, and c) contextual data (data used to check for important potential differences between treatment and control groups). This data may come from the same or separate sources. The sources need to be structured to form analytical datasets (or analytical samples) that are used to estimate impacts. This structuring will in many cases involve linking records of individual persons or enterprises across sources. Such linking requires either individual level identifiers (for example, individual social security identification numbers), that

enable an individual's record (for example in tax data) to be aligned with participation records, or enough data to link records across sources (for example, name and date of birth must be available across sources). It is important to consider which data sources will be exploited for the CIE being planned, but also whether it will be possible to link records across sources.

2.3.5. What are key constraints in analysing data and results?

As discussed above, impacts in CIEs are usually determined through comparing results in the treatment group with those in the control group and this fundamental comparison is, in essence, part of the CIE approach. The difference between treatment and control groups is referred to as the impact or net effect of the intervention. However, the precise way impacts are estimated will depend on the research design adopted.

In planning a CIE, it is important to consider whether the intervention is big enough and likely to generate impacts that are capable of being detected statistically. In fact, CIE methods are based on the principle that target and control groups are a sample of the whole potentially treated and control populations and, to produce significant results, the number of individuals included in the two groups has to be statistically significant.

Assessing sample size and effect size

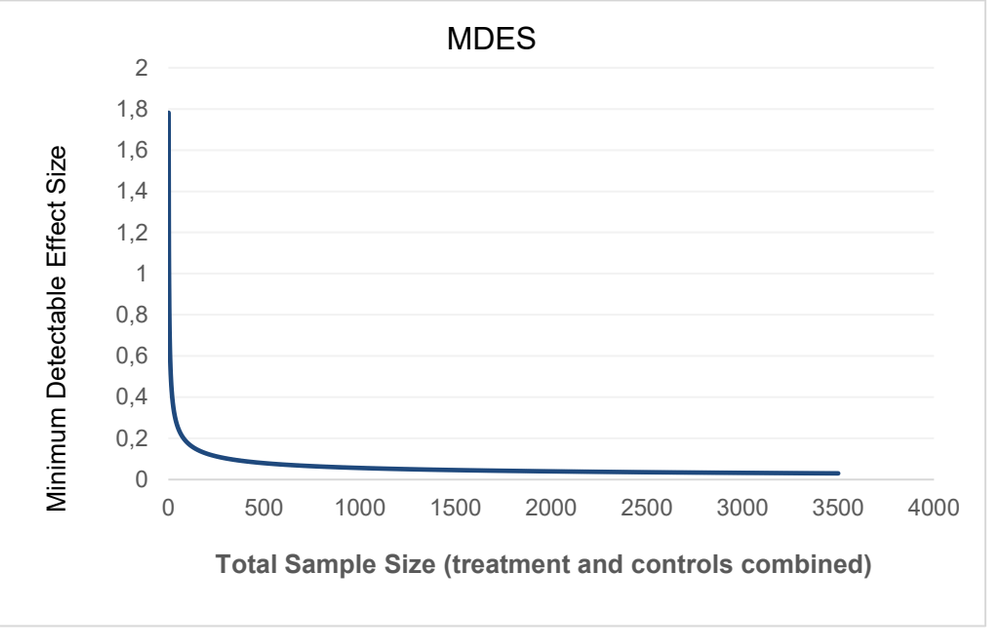
When considering whether a sample is of sufficient size for analysis, a useful concept is that of the 'minimum detectable effect'.[65] Simply put, a minimum detectable effect is the smallest true impact a sample size can detect at standard levels of statistical confidence and power[66].

Figure 4 below shows how the minimum detectable effect size varies with total sample size (total sample numbers in treatment and control groups) at a 95% level of statistical confidence and at an 80% level of statistical power. Moving from left to right, the minimum detectable effect size declines rapidly as the sample size approaches 500 (250 treatment units and 250 controls). In other words, as the total sample size increases, the CIE design is more precise and capable of detecting smaller impacts.

[65] Bloom H. S.,1995, provides practical guidance on how to calculate minimum detectable effects for experimental designs. In the case of quasi-experimental approaches, such calculations will require adjustment. Generally, quasi-experimental approaches require larger sample sizes relative to those necessary for an experimental design.

[66] The level of statistical confidence is a statistical measure of the reliability of the estimation procedure, while the level of statistical power is the likelihood of a test detecting a real effect. These two parameters also depend on the sample size and, inversely, they can be used to calculate a sample size in respect of adequate levels of significance. In general, statistical confidence is 95% and statistical power is 80%.

Figure 4 Minimum detectable effects sizes (MDES) at different sample sizes



Note: 95 percent statistically significant and 80 percent statistical power. A randomized design is assumed in the figure

The likely size of the sample

In planning a CIE, it is useful to estimate the likely size of the samples to define treated and control groups; this estimation is based on forecasts of the number of units that will be treated, the design of the probable method that will be used in the CIE and the size of corresponding control groups derived from available administrative data or a suitable survey. This information can then, under certain assumptions, be used to verify whether the resulting minimum detectable effects are sufficiently significant, and whether implementing the CIE is a reasonable choice. A key aspect in this exercise is whether the intervention concerned is likely to generate effects of a size equivalent to the estimated minimum detectable effects; on this point, an examination of the existing literature and other similar evaluations can help.

It is also noteworthy that CIEs often intend to examine the effects on different sub-populations involved in the intervention (male/female, young/old, less/more educated, etc.). These analyses necessarily reduce the size of the treated and control groups for each sub-population (male, female, young, etc.) and consequently the statistical significance decreases. When the study of the effects in the sub-populations is an important component of the evaluation, it is necessary to have a sufficient number of treated and control people in each sub-population.

The definition of the sample size requires technical skills which are not always present in the MA. In this case, the evaluator has to calculate the sample size and agree with the MA on the feasibility and the scope of the CIE. It is not possible to define the threshold sample size above which it is always possible to carry out a CIE, because the threshold can differ depending on the methods adopted, the expected analyses of sub-populations, the accepted levels of statistical significance, and other elements. However, in many evaluations a number of 2,000 individuals for the total of treated and control groups has proved to be sufficient for a reasonably accurate analysis (including main subgroups, such as male /female etc.). This indication cannot be used as a

scientific “threshold” but simply as an aid in examining the potential implementation of a CIE.

Sample size affects the robustness of CIE results. Some examples of the uncertainties in interpreting results and their relation to sample size are explained in the box below.

Box 12 Uncertainties in interpreting results

Among examples of evaluations of ESF-funded interventions, the evaluation of several measures for long-term unemployed in the Marche Region in Italy did not deliver statistically significant results in one of the four measures analysed, ‘Vouchers for Training’. This is probably due to the fact that the sample of treated people under analysis was small in comparison to the large number of actual recipients. A similar case occurred in the evaluation of measures to promote vocational education in Poland, Podlaskie Region. Apart from the issue of statistical significance, this example shows that, when analysing the effects, reliance on overly small samples of treatment groups can lead to problems of generalisation in the findings.

The evaluation of language training for immigrants in Germany shows positive results, in contrast with some previous studies assessing impacts of language training programmes for immigrants. A potential explanation of this difference, according to the author, is that the German programme under analysis also consisted of a work experience module. This hypothesis could not be tested by the author with the available data, because of the impossibility of distinguishing between the language training and work experience components of the intervention. This can be of potential interest for future research. The example shows that sometimes counterfactual exercises need to be repeated with improved and richer data or complemented by other evaluation approaches in order to explain the mechanisms which determine the results of an intervention.

2.3.6. A check-list to verify preparation and feasibility of the CIE

A list of key factors to take into consideration

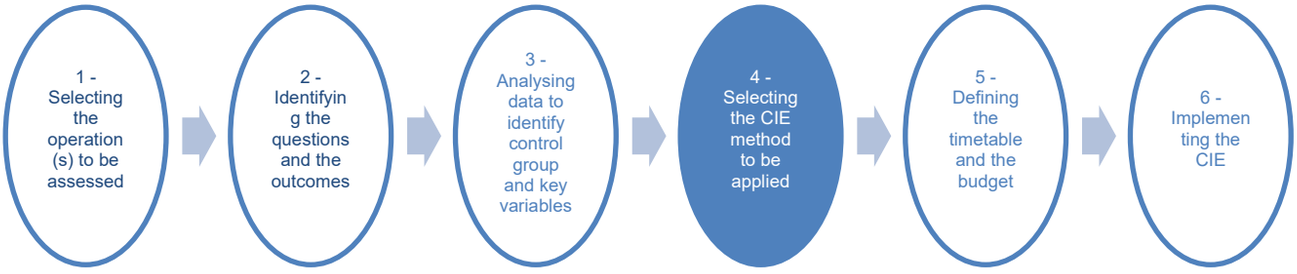
At this point, it is useful to summarise in a basic check-list the main factors that an MA should consider and verify in the preparation of a CIE:

- Is the selected intervention for evaluation suitable for a CIE?
    - Is the intervention discrete, distinctive and relatively homogenous?
    - Does the ‘theory of change’ of the intervention suggest a convincing causal mechanism of the outcomes to be examined by a CIE?
    - Is a control group easily identifiable in accordance with the rules of the intervention?
- Are the effects on the participants quantitatively measurable?
    - Are the outcomes of the intervention being measured clear and consistent with the theory of change of the intervention?
    - Has sufficient time passed since the end of the treatment to detect outcomes for the participants?
- Do evaluation questions require a measurement of the impact (net effects) of the intervention?
- Are appropriate data (treatment and control group records, outcome records, context data) available or can they be made available?
    - Does available data comply with GPDR? If not, can the necessary inter-administrative agreements and technical solutions be activated to make data compliant?

- Is it possible to define a proper control group with the available data?
- Is the sample size big enough to reach the required statistical significance within the CIE results?

An affirmative answer to all these questions means that it is possible to design and implement a CIE. A negative answer to some of these questions does not rule out the implementation of a CIE, but requires additional checks on the feasibility conditions and improvements in data availability or other key factors.

2.4. CIE method to be applied



Some fundamental conditions for identifying the most suitable method

Selecting the method to be used in the CIE is a crucial step in the preparation of the evaluation since the method selected affects both the quality of the findings and quality overall. In general, no method is superior to another, but one may be more suited to exploit available data while another may be more appropriate for investigating a certain type of intervention; identifying the most suitable method for a specific evaluation is a crucial step in the design.

The technical aspects and the intrinsic characteristics of each method are examined in detail in **Chapter 3**, here it is useful to reflect briefly on the practical implications of selection.

First, a randomised approach - that randomly assigns individuals to the treatment or control group - has to be established before the beginning of the intervention. This assumes both the completion of the evaluation design and an organisation capable of tackling the task. In ESF interventions this approach has rarely been adopted, but, as we will see, it is possible under certain conditions and very effective from the methodological point of view.

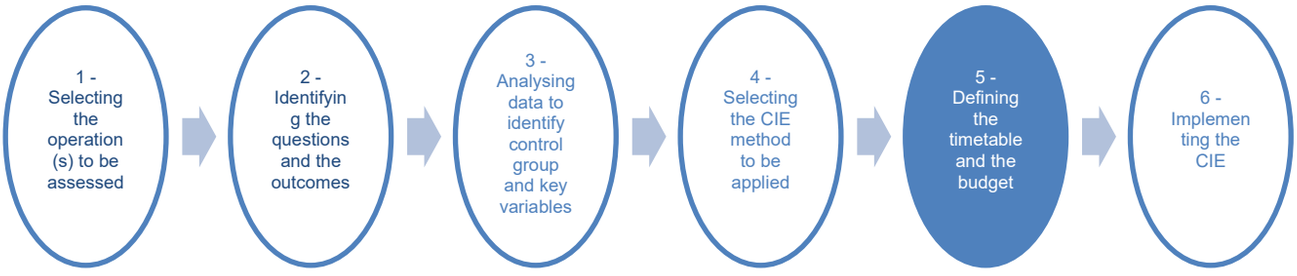
Second, the selection of the CIE method may be beyond the capacities of the MA, because it requires advanced technical skills and specific experience. In such a case, the evaluator will select the most suitable method, but, clearly, only after he has been nominated. This can have the added bonus of acting as a criterion to assess the capacity of evaluators in the procedure of evaluator selection. At all events, the MA must feel confident in carrying out a CIE after having given positive answers to all, or most of the questions in the previous check-list and has acquired – internally or externally - the necessary skillset to approve the choice of the evaluator.

Third, randomized methods can differ from quasi-experimental methods in terms of duration and costs and in randomized control trials the evaluator should be involved from the beginning of the intervention. Quasi-experimental methods do not differ significantly from each other; the evaluator could start

work after the intervention has begun (but in time to organise and prepare the necessary data) and the quality of the available data could make the difference in terms of time and costs.

Finally, the type of intervention and data availability sometimes allow the use of more than one quasi-experimental method. In this case, it may be useful to apply more methods because confirmation of the same effects using different methods improve the robustness of findings.

2.5. Timetable and budget



2.5.1. What resources are available?

A key issue to consider in devising a CIE evaluation scheme is resource availability. This can be a wide-ranging set of considerations, here arranged under three headings: a) expert resources; b) time; and c) financial resources.

Which external experts and internal staff are required for a CIE?

Internal personnel

In most cases, an impact evaluation will be contracted to an external supplier. However, the contract will need to be managed within the MA by staff with knowledge of CIE methods. Such knowledge is required in order to ensure quality and to liaise effectively with external experts. Other forms of expertise may also be required within the MA, such as statistical skills, and expertise in data collection and management. It is important to consider in advance whether the MA has access to suitably qualified and trained staff, and that these staff have the capacity to support the evaluation.

External personnel

Commissioning an effective CIE requires contractors who have the necessary skills and experience to conduct such evaluations. In addition, suitable contractors will need to understand the policy and administrative context within the MS, be familiar with potential data sources and be proficient in the appropriate languages. It is worth considering what is required in order to develop a CIE supplier base within an MS (for further discussion on this topic see Chapter 4).

Staff managing programmes / interventions

Effective CIEs require cooperation from those managing the programme or intervention being evaluated. For example, access to registers maintained by intervention managers will be required. These registers provide information about individuals or enterprises who participated in an intervention.

Programme/intervention managers can provide advice and guidance on these types of data. They may also be required to conduct record-keeping in addition to what they would need to do in the absence of an impact evaluation.

Statistical expertise

To overcome the issue of data collection from various sources, those planning a CIE will need to liaise with staff dealing with official data sources (e.g.,

unemployment registry, social security data, statistical offices, etc.). This will enable planning data provision well in advance.

Which factors are relevant for timetabling a CIE?

CIEs take time

Conducting a CIE requires contributions from a number of different human resources; in addition, such evaluations are carried out over considerable time spans. An evaluation scheme should contain an outline timetable with crucial project milestones, which will be valid for those involved with the intervention itself, as well as those associated with the evaluation. The outline timetable will need to be organised across both evaluation and intervention delivery activities, in addition to key policy milestones.

Planning a time lag to allow impacts to emerge

Developing a meaningful and realistic outline timetable for a CIE is a difficult balancing act. On the one hand, the managing authority (or IB) planning the evaluation needs to consider the key dates by which decisions that depend on the evaluation's findings will have to be made. On the other hand, there will be inevitable constraints, which will impinge on the timing of reports. Some results will take years to emerge, and data collection, analysis and reporting timetables will need to reflect this as far as possible (see Section 2.5.2 below).

Focus on specific points in time

The evaluation needs to be conducted early enough during the programming period to enable changes to be made and experiences and lessons learnt to be capitalized upon in the remaining time. In some circumstances, the same or similar interventions might be supported in successive programming periods. Results from CIEs focusing on interventions from previous programming periods can be extremely helpful in informing implementation and design in subsequent programming periods.

Sequence various types of evaluation

It is also important to consider how the timing of a counterfactual evaluation might relate to the timing of other evaluation components. It is likely that a theory-based evaluation would need to be completed prior to a CIE. For innovative interventions (e.g. 2014-2020 ESF interventions that supported children in the Czech Republic, or the services fostering social inclusion for individuals receiving the new income support in Italy, or the various measures to support workers and students in distance working during the COVID-19 lockdown in many MSs), key elements of a process evaluation may need to be reported either earlier, to enable improvements in the design of a CIE, or at a later point to facilitate a detailed investigation of causes and effects. When conducting a CIE of a mature ongoing intervention, it would be more relevant for the process evaluation to be conducted alongside the impact evaluation.

Data collection can be time-consuming

A timetable will also be affected by the availability of data. Data sources can take significant periods of time to update, as is often the case with tax records. Surmounting legal and institutional barriers to acquiring the requisite data can also be time-consuming and expensive. Moreover, drawing upon data from a range of sources, ensuring their compatibility, checking their quality and moulding them into a form that can be used to estimate impacts requires considerable time and effort.

How can costs be assessed?

It is important to set an indicative budget for how much an MA is able and willing to spend on a CIE. The budget is made up of two components: internal and external costs. The former refers to the effort required of the internal resources to follow the evaluation; within administrations these costs are

generally not detailed in financial terms but have to be taken into account to ensure adequate control and follow-up of the evaluation. Costs of commissioning external experts to conduct the CIE have to be carefully estimated to ensure a high-quality evaluation at reasonable expenditure. The focus here is on external costs.

A possible method for assessing costs

A distinction needs to be made between evaluations of routine interventions, where expenditure is generally lower, and innovative or pilot actions for which the collection of a relatively high amount of data, the use of new data sources and the involvement of new stakeholders may justify higher expenditure. However, this is not a rule and must be seen in the context of the evaluation questions, whose complexity and number may require lesser or greater financial input.

The choice of the evaluation approach also makes a difference. As mentioned above, a randomized approach requires the presence of the evaluator to collect and verify information for the duration of the intervention. On the other hand, the effort in quasi-experimental methods is largely determined by the number of data sources, their quality and availability.

A guidance document issued by the Commission[67] provides an indication of the sum needed for the evaluation of a programme and states that large scale and routine programmes should dedicate no more than 1% of their programme budget to evaluation; while innovative or pilot initiatives may commit up to 10%. This guidance, however, does not explicitly address the resource needs of CIEs and we can assume those percentages to be the ceiling for evaluation costs including one or more CIEs.

As one can imagine, it is impossible to provide an indicative cost of a CIE for such a huge number of different interventions and operational conditions in ESF+ programmes in 27 MSs. Here, it is more useful to suggest a method for estimating a reasonable budget for CIEs in different contexts. The following table illustrates a simple approach: the rows list the main tasks to be carried out, while the columns report the main costs.

The duration and the complexity of each activity determines the effort required in terms of working days of the evaluation team and, consequently, its main cost. Costs of data collection in the form of quantitative surveys of participants and control group members are considerable. Where a CIE relies on existing administrative data sources and these are quite easily accessible, total costs will be lower. However, administrative data often needs essential ‘preparation’ activity (correcting missing or wrong data, adjusting the format of the database, verifying the administrative rules governing data, etc.) which can increase costs.

[67] See European Commission, 2013.

Table 3 Structure of the main costs of a CIE

Main Activities	Evaluation team			Other Costs	Notes
	Program Manager	Senior experts	Junior experts	(Equipment, materials, supplies, travel, etc.)	
Planning and coordination					The planning process includes the feasibility of the CIE, its organisation and the finalisation of the methodology in agreement with the MA. In this phase sufficient time needs to be dedicated to the analysis of potential data gaps. The coordination covers the entire duration of the CIE and includes organisation of work and interactions with the MA and other relevant stakeholders to finalise the evaluation questions.
Literature review					The literature review supports the outline of the theory of change, identification of variables to be used and understanding of the context. It also makes it possible to refine the evaluation questions and exploit past results and formulate hypotheses.
Data collection and preparation					The effort required for this activity varies widely depending on the data collection method(s). Original data collection, via surveys, can be time-consuming and expensive. Other methods may require investment in technology (software or hardware) or reaching agreements with data owners. Costs can be reduced by involving data owners in data preparation.
Data Analysis					Data analysis requires advanced skills and experience. Time taken for this activity is influenced by method, quantity of analyses and data quality.
Report(s) Preparation					The effort for this activity varies based on the number and type of reports and other communication tools. Costs can include printing and graphic design in addition to preparation time.
Follow-Up Meetings					Follow-up meetings and other dissemination activities are an important step to disseminate findings. Miscellaneous costs could include space rental and food.
Total					

In the absence of other references, the cost of staff can be calculated in relation to ESF+ costs for training or employment services adopted in each country for senior and junior trainers or experts. These fees are generally comparable to those of a junior and senior researcher and, even if they do not necessarily have to be equivalent, are a useful parameter to take into consideration. A simple market analysis and some interviews with a few researchers may help to specify the expected cost of the CIE.

2.5.2. When should the intervention be evaluated?

It is crucial to determine when in the life of an intervention it is most appropriate to conduct an impact evaluation, as well as the critical issues of when results should be measured and impacts estimated.

When to evaluate new and ongoing interventions?

Timing differs for new and ongoing interventions

Discussion of when in the life of an intervention it is appropriate to conduct a counterfactual impact evaluation will be shaped by whether the intervention is new or a mature ongoing scheme. For a new intervention more time is needed to become mature and reach a stable state. Conducting a CIE before this point will be premature and potentially provide misleading evidence. In the

case of new interventions, an initial process evaluation, conducted prior to a CIE, is often useful to identify teething problems and indicate solutions.

Resource issues

In determining the optimal timing of a CIE for a new intervention, a range of other factors should be considered, these include: steps to ensure that appropriate data sources are available, establishment of an internal project team comprised of appropriately trained personnel and appointment of an external contractor. Furthermore, the needs of the decision-making process at which the evaluation is ultimately directed will be a critical constraint.

The timing of an impact evaluation for an ongoing intervention will be driven mainly by practical and policy-related requirements. The intervention should already have bedded down and reached a level of maturity, making a CIE appropriate. One further issue that should be considered is the presence of other reforms running parallel to the intervention being evaluated. The effects of these reforms may influence the impact of the intervention being considered. Policy makers will need to consider whether the presence of other reforms on the policy landscape is relevant to policy decisions which will draw on the results of the CIE under consideration.

Multiple programming periods

ESF evaluations are usually focused on one programming period. However, especially in the case of stable interventions already part of the ESF programme in the previous period, it may be worthwhile combining a retrospective evaluation of the previous period with an ongoing evaluation of the current period in order to cover a longer life-span of an intervention.

When to measure results and calculate impacts

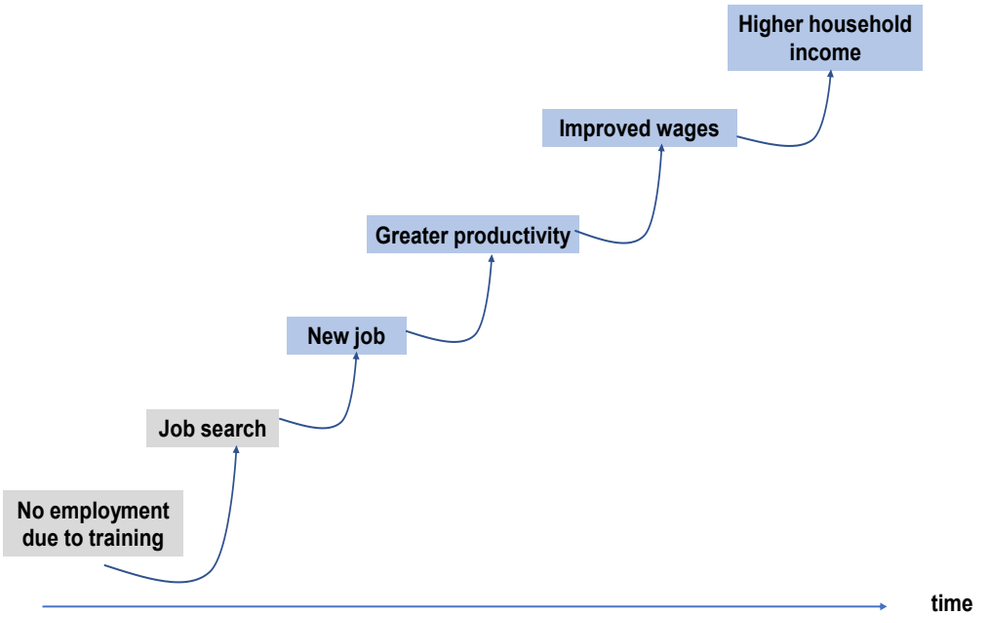
The second key issue associated with the timing of an evaluation is when impacts should be measured and estimated or, more specifically, when it is anticipated that impacts will emerge following an intervention.

When examining a training intervention targeted at the unemployed, the question is over what period of time higher rates of employment might be expected. It is a well-established feature of training programmes that in the short run they tend to reduce employment among participants. This is due to the so-called ‘lock-in’ effect. Training interventions tend to divert unemployed trainees away from job searches due to their attendance at courses. In contrast, control individuals are committed to finding a job. If impacts are calculated too soon, they may well be negative or underestimated. In planning a CIE, it is important to be realistic about the timing of impacts and when they are likely to be measurable. A simplified model of subsequent impacts is given in Figure 5 below.

Between the reasonable and the feasible

Consideration of when best to measure results and estimate impacts will need to take account of policy makers’ requirements for information by certain deadlines. In the case of interventions that aim at improving long-term employability, it may make sense from an analytical perspective to follow up participants two years or more after they have been exposed to treatment, to see if their earnings and rates of employment are higher than some equivalent group of untreated persons. In contrast, programme managers often need findings quickly, in which case a compromise has to be reached between what is a reasonable follow-up interval from the perspective of the intervention and decision makers’ need for timely evidence.

Figure 5 Simplified timeline for results of a training programme



Focusing on specific points in time

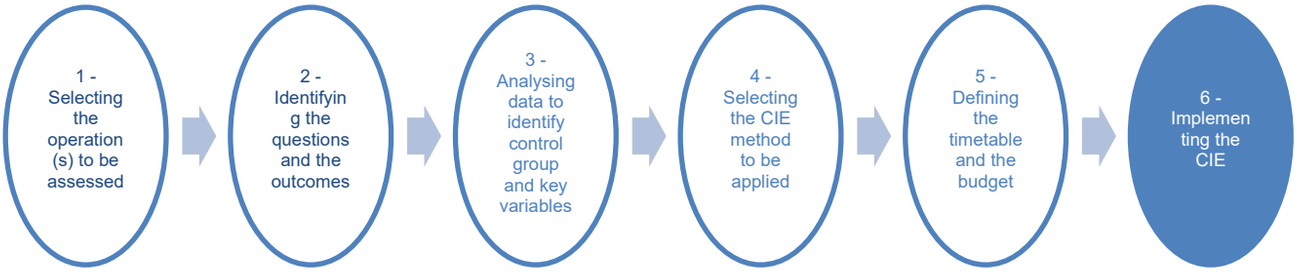
If result measures are obtained from administrative sources (e.g., social insurance records with details of employment and earnings), then it will be practical to track results repeatedly over a sustained period of time and estimate impacts (even on a monthly basis). The risk here is that the nature of findings may change over time. If primary data collection is required for the measurement of results in the form of sample surveys, estimating impacts at regular intervals would become very expensive, unless retrospective data on results can be viably collected. However, the cost of extracting data from multiple administrative systems and creating a single analytical dataset should not be underestimated.

As discussed in Section 1.5, the articulation of a theory of change (or intervention logic) can help determine when to estimate impacts.

...or reviews of other recent studies

An alternative for those planning a CIE in the absence of a theory of change (but also useful for those who can draw on a clear theory of change) is to conduct a short review of previous studies evaluating interventions which are similar to the one being considered. Careful consideration of results from previous studies can give a good indication of the appropriate measurement of results and calculation of impacts.

2.6. Implementation of the CIE



With the definition of the timetable and the budget, the planning of the CIE is completed and it is possible to move to its implementation. The final step in this Guide intends to illustrate key activities in carrying out the evaluation, and how the MA can overcome potential obstacles confronting them. In particular, this section applies the lessons from the Evaluation Helpdesk, through which the EC and DG Employment provide assistance to MAs of all MSs. Implementation of the evaluation is discussed in general in "EVALSED: The Resource for the Evaluation of Socio-Economic Development"[68]; here the Guide considers four main activities, specifically from the point of view of CIE:

1. Selecting the evaluator,
2. Supervising the implementation of the CIE,
3. Reporting,
4. Using the results.

2.6.1. Selecting the evaluator

The independence of the evaluator

The first decision when selecting an evaluator is whether to use an internal or an external candidate. In the latter case, the selection of the evaluator takes place through a public procurement procedure.

In art. 44(3) the CPR recognises the existence of an independence issue and states that: "*Evaluations shall be entrusted to internal or external experts who are functionally independent*". Hence, the first criterion in selecting an evaluator is their genuine independence from the MA and the decision-making processes of the ESF+ programme. The second criterion is whether or not the necessary skills and resources to carry out a CIE are internally available. This is probably the most frequent obstacle to internal evaluations, because technical knowledge to conduct a CIE is very specific and rarely available in most administrations.

Main elements of the terms of reference

Once the choice between internal or external evaluator has been made, the administration has to list the technical specifications of the evaluation; in the selection of an external evaluator the specifications will be part of the public procurement procedure. In the terms of reference for a CIE the following aspects deserve particular attention:

- The **aim of the evaluation** and the research questions must be clear and consistently lead to an impact evaluation and a CIE, where possible;
- **Available data** must be specified; this is one of the main issues to tackle in designing and implementing a CIE. It makes evaluators aware of the

[68] Available here: https://ec.europa.eu/regional_policy/sources/docgener/evaluation/guide/guide_evalsed.pdf

existing problems in data availability and allows them to draw up realistic and accurate proposals;

- **The request to conduct a CIE** can be explicit in the terms of reference, but the nomination of specific methods to be used should be left to the evaluator. In this way they can propose original solutions and demonstrate their capacity.
- **The proposed team** has to combine expertise in several fields, particularly in: evaluation and CIE (knowing how to design and implement an evaluation), econometrics (using the right methodologies), and the policy field of the assessed intervention (reconstructing the theory of change and interpreting findings in full).
- **The award criteria** have to prioritise quality over the cost of the proposal. The quality of an evaluation largely depends on the evaluator's ability to design and execute it. Legal frameworks for public procurement are often not a perfect fit with the needs of evaluation tenders, but the commissioner should adapt, as far as possible, the existing rules to reward quality and technical capacity.
- **The selection process** should include experts in evaluation and CIE on the selection committee in order to fully appreciate the quality of the proposals and correctly assess their various methodological and organisational solutions.

In some MSs, experience with CIE is still negligible. This may create a problem in the selection process because there the market for this service is too restricted. Possibly very few companies or experts will be able to comply with the technical conditions of the terms of reference and deliver a good quality proposal. In these cases, open seminars of the MA with academics and companies interested in evaluations may help to prepare the CIE and the subsequent call for proposal. The seminars would collect suggestions from national experts and give them the time to organize participation in the call for tender; to respect transparency and avoid conflict of interest, seminars should be public, and basic information should also be made available to those who could not attend.

2.6.2. Supervising the CIE

After choosing the evaluator and the kick-off of the CIE, the MA has to supervise the evaluation process. This mainly involves controlling and validating the evaluator's deliverables, organising progress meetings with the evaluator and stakeholders and, if necessary, reaching agreements with data owners.

Internal and external skills in the supervising process

The MA must have the capacity to fulfil these tasks. This also means having resources to supervise the entire process and skills to check technical solutions of the CIE. In general, the MA dedicates specific staff to the evaluations and, where they have insufficient experience with CIEs, independent external experts should give support. These experts may be taken on with short-term contracts or, when they come from public institutions, services will be free of charge. The commitment required will probably involve between 6 and 12 working days, which should suffice to review reports and attend some meetings with the MA, evaluator and main stakeholders.

Setting up a steering group is a frequent way of supervising the evaluation processes of a programme. This group is generally comprised of the MA,

officials representing other implementing Departments, main stakeholders (social partners, NGOs, other administrations) and some experts from academia and/or public institutions (e.g., the national statistical institute). The steering group participates in the definition of the evaluation questions, discusses the evaluation findings, their dissemination and use in policy making processes. A more flexible and specialised sub-group could also supervise the evaluation deliverables with the MA.

2.6.3. Reporting

Reporting is a key activity of the evaluation and is the principal means of communicating results. Evaluation reports must be clear, concise, intelligible for non-technicians and transparent in their judgements and policy recommendations. They must also demonstrate the reliability of the evaluation by specifying the methodology and data used, as well as identifying the possible limits of the analysis.

In general, there will be three reports for each CIE:

Distinct reports to follow progresses and use results

- **Inception report** is usually delivered shortly after the signature of the contract and presents the finalisation of the methodology in accordance with discussions held with the MA and the initial screening of the available data.
- **Interim report** is delivered at an intermediate stage of the evaluation and in a CIE may be devoted to showing data collected, the composition of treatment and control groups, and sampling.
- **Final report** presents the analyses and findings of the CIE. It includes details on methodology and data, and highlights any policy implications emerging from the evidence. The structure of the final report of a CIE must include some key sections: evaluation questions; theory of change (or intervention logic) and identification of outcomes to be assessed; adopted methodology; data used and characteristics of treatment and control groups; estimated effects; answers to the evaluation questions and their policy implications.

The commissioner reviews all the reports with the support of external experts where necessary, and comments are discussed with the evaluator, who will process the requested clarifications and improvements prior to the formal acceptance of the report.

A CIE report must avoid excessive technicalities and make findings clear and readable to stakeholders and non-experts. Specific annexes may include the more technical steps of the analysis, ensuring a plain and fluid narrative in the main report. However, methodological details are crucial to demonstrate the reliability of the CIE and must also be transparent. The identification of outcome variables, data quality, composition of treatment and control groups and sample size, the estimation method of the effects and the statistical tests used must all be included in the report. Furthermore, in line with academic practices, the anonymised data used and the methods of data processing could be made accessible to researchers who intend to repeat and verify the analyses.

A minimum set of information to be provided in the final report of a CIE

A standardized set of minimum information should be provided in the final report of all CIEs. This minimum set of information has a twofold objective: 1) to facilitate a comparison of findings across different OPs and different

countries; 2) to collect information which will be available for future meta-analysis.

The template for collecting a standardized set of minimum information on each ESF evaluation follows the recommendations spelled out in the "Pilot and feasibility study […]". From an operational point of view, this information could be included in an annex to the final report and could also be sent to the EC.

Table 4 below presents a possible template for collecting the minimum set of information needed.

Table 4 Basic information to include in a fiche presenting the CIE

Information category	Specific information to be collected and reported
1. **Intervention information**	- Name of the evaluated intervention - Characteristics of the intervention (training, employment services, services for social inclusion, etc.) - ESF+ programme approach (priority axes, specific objective, action) - Managing authority and/or other implementing bodies - Expenditure - Duration/intensity of the treatment provided by the intervention - Participant group by age, gender, and eligibility status (UI recipient, Long-term unemployed, NEET) - Territorial scope of the intervention
2. **Measure of Intervention effectiveness**	- Indicator on whether the CIE estimates a positive and statistically significant treatment effect on the relevant outcome(s) - The actual size of that effect - The two above measures (+/-) and (size) for the set of relevant and comparable outcomes: e.g., employment rate, earnings, etc. - 'Value for money' indicators (if estimated)
3. **Data and methodology**	- Data source (survey, administrative data) - Observation period - Econometric CIE method - Time horizon (generally, short<=12 months from intervention end, medium>12 and <=24 months, long-term effects>24 months) - Sample size of treatment and control groups - Availability / inclusion of pre-programme data

2.6.4. Using the results

Disseminating the findings of the CIE is crucial

The use of the result is important because, unless results are disseminated effectively and reach their intended audience, the evaluation will have little impact.

Dissemination of findings and evaluation outputs usually involve:

- At least one written evaluation report, including an abstract and an executive summary;
- At least one verbal presentation of findings, supported by slides or similar tools;
- A technical section (or annex) of the report providing a thorough account of the methodology deployed, key assumptions made and the approach to statistical analyses adopted.

All evaluation reports need to be made public. This is a stipulation in the Common Provisions Regulation for the programming period 2021-2027[69]. In addition, evaluation findings have to be presented and discussed in the Monitoring Committee of the programme; this is a compulsory passage but is not sufficient in itself to promote an in-depth debate, because operational priorities generally prevail in the meetings of the Monitoring Committees. See the box below for an example of a national practice in presenting and disseminating evaluation findings.

Many audiences and many communication approaches

There is no single effective communication policy to disseminate and open a debate on lessons; each context requires its own strategy. It is, therefore, important to deliver an effective strategy, and especially ensure that stakeholders beyond the MA learn about the findings. It is possible to divide the "audience" of the CIE into main groups and identify the main types of communication; for instance:

- High-level policy makers and political representatives – they are used to seeing short documents (policy briefs, executive summaries) with main results and recommendations;
- Officials of the administration involved in the implementation of the intervention or similar policies – seminars or workshops can be an effective instrument to present CIE results, adapting the technical level of the presentation to the level of interaction with the attendants;
- Policy stakeholders (social partners, NGOs, beneficiaries, etc.) – summary of the reports and annual meetings on evaluation findings may be the combined instrument to involve these actors in the debate. In particular, when these actors have contributed to the definition of the evaluation questions, the reporting of the results of the evaluation should be linked to those questions.
- Experts and academics –scientific conferences or seminars may be the best venues to present the results of the evaluation. These actors validate the results of the evaluation from a scientific point of view and so reinforce their reliability.

Meta-evaluation to reinforce dissemination and generalisation

Meta-evaluation[70] is a powerful additional instrument for disseminating results of CIEs. It enables confirmation and generalization of the main findings that a single CIE cannot produce alone. To carry out a meta-evaluation it is necessary to have a good number of reliable and good quality CIEs on comparable policies at one's disposal; at the moment, meta-evaluations are still rare and not solely focused on ESF interventions. A mainstream use of meta-evaluations would require a more widespread use of CIEs and their anticipated planning in order to make their subsequent comparison easier. In the 2021-2027 ESF+ programming period meta-evaluations may be implemented in countries with a significant number of programmes or through cross-country comparisons.

[69] CPR 2021/1060 Art 44(7).

[70] That is a systematic comparison, even with the support of statistical tools, of numerous CIEs of similar interventions. A meta-evaluation provides indications on the effectiveness of a type of intervention by examining the effects of similar measures in many different contexts. However, meta-evaluations cannot enter into the detail of each single intervention used in the comparison, and their policy implications relate to broad policy orientations and not to single details for each intervention.

Box 13 The Polish experience with Evaluation Conferences

The International Evaluation Conference, organised by the Ministry of Development Funds and Regional Policy and the Polish Agency for Enterprise Development, is one of the key elements in the landscape of public intervention evaluation in Poland. Since its inception, it has become an established platform for discussion between stakeholders responsible for policy-making, policy implementation and effect assessment.

Started in 2005, the conference was held on an annual basis for 10 years, and bi-annually since 2015, with the 2021 event being its 14th edition. Currently, the conference is organised as a two-day event. One day is dedicated to research findings and effects of the interventions; it is a day on which policy is discussed. The second day focuses on methodological issues and evaluation as a process.

Throughout the 14 editions of the conference, counterfactual methods have been the subject of panel discussions or experts' presentations a number of times, and approached from different perspectives on both days. For example, in 2017 estimated effects were presented in sessions regarding the impact of the interventions, while discussions on advantages and disadvantages of estimation techniques were held separately.

The conference is also of note because of its diverse audience, since it addresses not only researchers or specialists in methodology, but also those who design and implement interventions. The goals of the event are twofold: building capacity of the evaluation system by developing the skills and knowledge of its members, as well as networking by providing stakeholders with a space to exchange experiences between national, regional and foreign partners. In fact, each edition brings together some 300-400 representatives from various sectors – public administration, academia, consulting companies and NGOs. It is a considerable number, given the fact that the evidence-based approach to policy has not yet fully taken root in Poland.

The International Evaluation Conference forms a part of the public administration evaluation system. Significant capacity building was possible thanks to training provided by the Joint Research Centre, the World Bank, or funded under Technical Assistance in operational programmes. Training on evaluation, including the counterfactual approach, was available not only to employees of evaluation units at Management Authorities, but also to institutions such as local or regional employment offices. The conference is an opportunity to gather, discuss and share current developments and findings[71].

[71] Information about the conference can be found in: https://www.ewaluacja.gov.pl/strony/xiv-miedzynarodowa-konferencja-ewaluacyjna/. We would like to thank Piotr Strzęboszewski and other officials of the Ministry of Development Funds and Regional Policy for the information provided on the conferences system.

Chapter 3. How to select the appropriate methodology to carry out a CIE

This chapter presents empirical methods for Counterfactual Impact Evaluation. Specifically, the experimental approach (randomised controlled trial) is discussed, as well as the most common quasi-experimental methods: matching (on the propensity score), difference-in-differences, regression discontinuity design, and instrumental variables. Each method is characterised by a specific way of generating a control group to answer the counterfactual question "What would have happened to the treatment group in terms of outcomes if it had not participated in / been exposed to the intervention?". In practice, the chosen design can be tailored to the given context, as determined by the type of intervention and the data that are available or can be collected[72].

Selecting the right approach

It is not possible to provide detailed guidance on the choice of the most appropriate evaluation design across what are highly varied circumstances faced by MAs. When choosing the most relevant approach to CIE in a particular set of circumstances, MAs should consider what has worked well in previous evaluations, both within the MA itself, within the MS and in other MSs - MAs can learn from what has been achieved before within their programme and elsewhere, where similar circumstances have applied. Forums for the exchange of lessons learnt in evaluation design and implementation can be useful sources of information in this regard. Searching the literature for evaluations of similar interventions can also be an important source of information to aid in the design process. Experts commissioned by the MA will also have views as to how best to approach an evaluation design. It is important to remember that there may be considerable expertise and experience within MAs that can be drawn upon.

3.1. Randomisation - the experimental approach

Key features of randomisation

- Individuals eligible for participation are randomly allocated to a treatment or a control group
- Randomisation ensures that both groups are identical (on average) in all relevant characteristics
- Hence, the control group answers the counterfactual question and the difference in outcomes between treatment and control groups indicates the causal effect of the intervention

Randomised evaluation designs can take any of a number of forms. Here the focus is on a straightforward two-group approach – one treatment group and

[72] For other technical methodological aspects see also European Commission, 2019.

one control group – in order to clarify the key principles. Figure 1 illustrates a simple randomised design.

Statistically equivalent groups

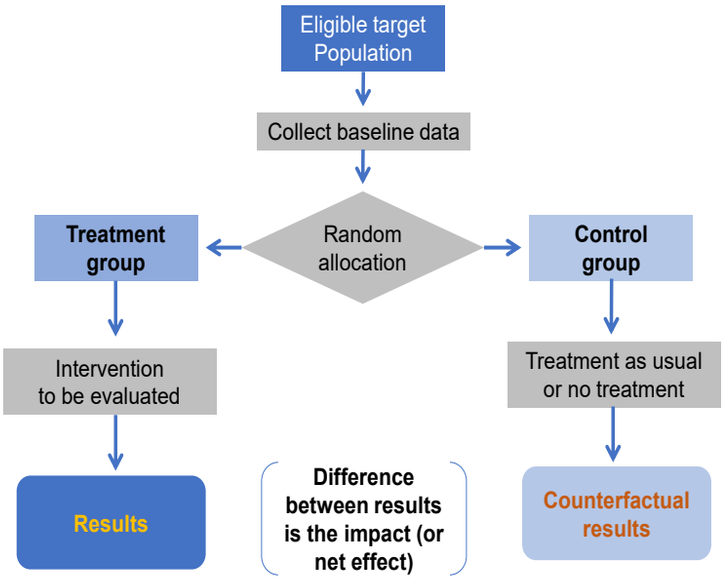
The key point is that randomisation ensures the two groups are statistically equivalent in all respects at the time of randomisation. Subsequently, the treatment group is exposed to the intervention which is the focus of the evaluation and whose impacts or effects are to be measured.

No treatment or other treatment for control group

Depending on the policy question of central concern, the control group can be assigned to receive no treatment at all, or the treatment group can be compared to a group exposed to some other treatment of interest (may be conceived as representing treatment as usual), or there can be multiple treatment groups alongside a control group. For example, there may be interest in comparing the effects of an ESF-financed training programme to other nationally-financed training, or to the provision of other services to the same groups.

Since treatment and control groups are statistically equivalent at randomisation, and exposure to subsequent treatments is controlled, differences in results can be attributed to the intervention being evaluated (subject to standard statistical uncertainties), and alternative explanations ruled out as the causes of any observed differences (see box below for an example).

Figure 6 Two-group randomised control



Strong evidence…

As a result of their intrinsic design features and if implemented correctly, randomised designs offer the prospect of providing strong evidence of an intervention's effects. They are highly favoured for this reason. However, they require early and detailed planning and can be intricate to design and administer. Furthermore, programme managers face significant challenges in implementing them correctly. For instance, the presence of the randomisation process itself may alter the composition of those who take part in an intervention: that is, some potential participants may be put off by the idea of randomisation and refuse to participate. Furthermore, individuals subject to

randomisation may not always comply with their assignment status, and there are a range of other challenges that may need to be confronted. In some circumstances randomised control trial designs can be expensive to implement.

… but difficult to design

For these and other reasons, it seems unlikely that evaluations of ESF-financed instruments and interventions will be conducted using a randomised approach. However, this Guide cautions against the impulse to rule randomisation out in all cases without proper consideration. The approach has been widely used and examples additional to that from the UK discussed in the following box include the GAIN experiments from the United States [73] (there are numerous other examples from North America), experiments conducted in Sweden[74] in 2014-2020 ESF interventions (see the next box), as well as a study undertaken in Germany to assess the effects of active labour market services supplied by private providers compared to those supplied through the public employment service[75], among many others.

Box 14 An example of a randomised trial of an ESF project for young people

The evaluation of 'Ung framtid' (Young future) project in Sweden

The Swedish 'Ung framtid' (Young future) is an ESF-funded project implemented between 2015 and 2018. It was aimed at supporting young jobseekers aged 18-24 entering the labour market in Middle Norrland, North Middle Sweden and South Sweden. The project, Arbetsformedlingen, intensified and personalised support activities targeting young jobseekers and was carried out by the Swedish Public Employment Services (PESs) –; it consisted of individual planning, information, advice and concrete support in terms of matching, training and other possible activities. Overall, almost 17,000 young people were reached by the project.

The evaluation was commissioned by the Swedish ESF Council to the Arbetsförmedlingen, as part of the evaluation framework developed in a 2014-2020 ESF project aimed at improving the evaluative capacity of the PESs (Evidence-based EU fund Project 2014-2020; see box in the final chapter of the Guide).

In an experiment conducted between June 2017 and January 2018 involving eight local employment offices out of a total of 90 participating in the project, eligible young people were randomly allocated to the treatment group (people received intensified support) and control groups (people received regular support)[76]. 4,689 young people were randomly allocated to treatment and control groups in the experiment, with 2,972 of them allocated to the control group. The random allocation process produced treatment and control groups that were very similar at the time of allocation. As a result, any differences between the two groups on the outcome variables (exit from unemployment, proportion of unemployed and average number of days in unemployment) measured after entering the ESF intervention, is attributable to the Young Future activities.

Findings from the study show that the Young Future project had positive impacts on women but not on men. For women, positive effects were found in all outcome variables considered: the exit from unemployment of treated women was higher by about 7 percentage points in the first two months after entry into the project, the proportion of unemployed decreased more in the treated than in the control group, and, finally, treated women were unemployed for fewer days than women in the control group[77].

Randomisation through research design

Randomised designs can be distinguished from other approaches mainly through their strong emphasis on controlling potential bias between treatment and control groups through research design. This emphasis on

[73] See Riccio J., Friedlander, D., Freedman S., 1994.
[74] See Hagglund, P., 2006.
[75] See Krug G., Stephan G., 2011.
[76] A pilot phase was conducted between April and May 2017 to test several activities, such as the management of randomisation, involvement of young participants, reporting.
[77] Axdorph E, Egebark J., Lundström T., Özcan G., 2019.

design makes the approach very intuitive, but requires advance planning. Randomised designs are often best implemented in evaluating new pilot interventions rather than existing ones. This is because they require some degree of control over how participants are recruited into the intervention being evaluated. This 'control' is often harder to achieve in existing programmes than in new interventions that are open to new ideas.

Ethical objections

Implementing a randomised controlled design means that a part of the eligible target population is not offered participation but is assigned to a control group. This assignment is entirely random and is not at the behest of either the applicant or the intervention's administrators. For this reason, policy makers may tend to object to RCTs on ethical grounds before considering whether they are feasible from practical and analytical perspectives.

...but also strong arguments for using randomised design

However, there is a strong case to be made for randomised designs. If randomisation provides the best quality, most reliable evidence of the effectiveness of publicly funded interventions, then it is important that they are used more widely in assessing intervention impacts. Further still, if the impacts of a certain intervention are a priori unknown, it is not unethical to exclude individuals, as we cannot assume that they would have benefitted. Moreover, such approaches are widely used in medicine and other fields of study such as education research. Finally, in some circumstances where the services and support provided by an intervention are over-subscribed (i.e., there are more eligible individuals than can actually be served by the intervention), assigning individuals to the intervention at random from the pool of those who qualify may be the most ethical means of allocating scarce resources.

3.2. Non-randomised or quasi-experimental designs

When randomisation is not possible

If randomisation is not feasible, a set of alternative methods exists in the CIE toolbox that use different approaches for generating a control group to answer the counterfactual question. This subsection first outlines the main challenge for quasi-experimental methods in general, then presents four approaches that are commonly used in empirical CIE practice.

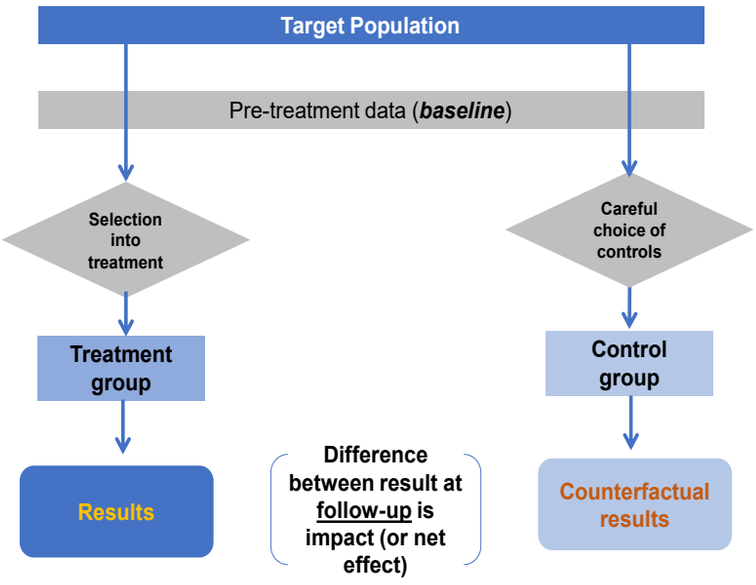
3.2.1. Target and control groups without randomisation

In quasi-experimental designs, target groups receiving the intervention are compared to a control group[78] of non-randomly allocated targets or potential targets that do not receive the intervention. As with an experiment, the objective is to obtain an unbiased estimate of the change in outcomes that the intervention under consideration has brought about. As treatment and control groups are not formed at random, quasi-experimental designs require far more attention to methods accounting for potential differences between treatment group members and potential controls which are likely to affect the decision to participate and therefore the results. The key is the selection of a plausible control group. Failure to select an adequate control group and account for remaining differences between the two groups in the

[78] Strictly speaking, "control group" is the term used for experimental designs, and quasi-experimental designs typically speak of "comparison groups". In practice, however, the terms are often used interchangeably. This Guide, therefore, uses the term "control group" for the group used to estimate the counterfactual, independent of the CIE design.

analysis weakens the credibility of estimates and can confound attempts to rule out alternative explanations for any observed effects.

Figure 7 Stylised quasi-experimental design with treatment and control groups



Control and treatment groups need to be similar to each other

In terms of ESF co-financed interventions, the most frequently applicable quasi-experimental evaluation design is a two group, baseline/follow-up design. Such designs feature a control group and a treatment group as in randomisation, except that the control group is constructed (without use of randomisation) from existing non-participant groups, making it as similar as possible to the treated group.

A good strategy for finding a valid control group within a quasi-experimental setting is to select control individuals that have been excluded from the treatment on the basis of factors unrelated to their characteristics and potential results. In some circumstances there may be reason to believe that although control groups have not been constructed explicitly at random, individuals or enterprises can be identified ex-post whose non-exposure to the treatment turns out to be random with respect to potential results. If these circumstances occur, they are close to ideal within the context of a quasi-experimental approach. For example, certain members of an intervention's target group may be excluded from participation in the intervention as a result of administrative oversight or error. Understanding the process of selection into the treatment is therefore extremely important in drawing up a valid control group - this cannot be emphasised enough.

Matching treated and untreated individuals

A credible control group can be developed in a number of ways. First, a statistical matching approach can be taken: i.e., data are collected from both treated individuals and a (typically very large) sample of non-treated persons. A control group is then constructed from the group of non-treated individuals by choosing those who are most similar to the individuals in the treatment group. "Similarity" refers to a set of socio-demographic characteristics – e.g., age, gender, education, employment and unemployment, etc.– measured at a point in time before the treatment group entered the programme. As a result,

non-treated individuals are effectively "matched" to treated individuals. In practice, the potentially long list of socio-demographic characteristics can be summarized using the value of the 'propensity score', which facilitates the application.

3.2.2. Propensity score matching

Key features of propensity score matching

- Matching mimics randomisation by constructing ex-post a control group that is as similar as possible to the treatment group in all relevant characteristics
- Unlike randomisation, only observable characteristics (age, gender, education, etc.) can be matched, while unobservable characteristics (e.g., motivation) cannot be taken into account
- The validity of the approach is heavily dependent on data availability

Propensity to participate as a way to define treated and control groups

Propensity score matching (PSM) entails estimating a statistical model for the entire sample (treatment and potential controls) which yields an estimated propensity to participate for each individual or firm - regardless of whether they actually participate or not.[79] Treated individuals or firms are then matched - either to one untreated individual or firm, or to several untreated individuals or firms - on the basis of the propensity score.[80] This procedure identifies a control group that can subsequently be used to derive an estimate of the counterfactual. Matching in this way ensures that impact estimates take account of the observable differences between the treated group and those acting as controls, and, assuming that all relevant pre-matching differences are observable, an unbiased estimate of intervention effects can be obtained. However, if selection into treatment is based on unobserved factors there will remain a question mark as to the adequacy of matching in terms of its capacity to eliminate bias. The critical assumption underlying the matching approach is that the selection process is characterised by observable data.

The figure below presents an intuitive and simplified illustration of the propensity score matching approach. The Y axis represents the number of individuals in the treated and non-treated groups, ordered by the propensity score on the X axis. Typically, treated individuals tend to have relatively higher propensity scores, while non-treated individuals tend to have lower propensity scores. The area in which the propensity scores for the two groups overlap is known as the region of common support.[81] Treated cases are matched to untreated cases within this region. Two examples are given in the diagram, but the process is essentially repeated until every treated case is matched to an untreated case within the region of common support. In the figure this is done using 'nearest-neighbour' matching. The 'nearest neighbour' to any member of the treatment group is the control group member with the closest propensity score. Once two groups have been formed, their mean results can be compared in order to obtain an estimate of impact. In practice, carrying out

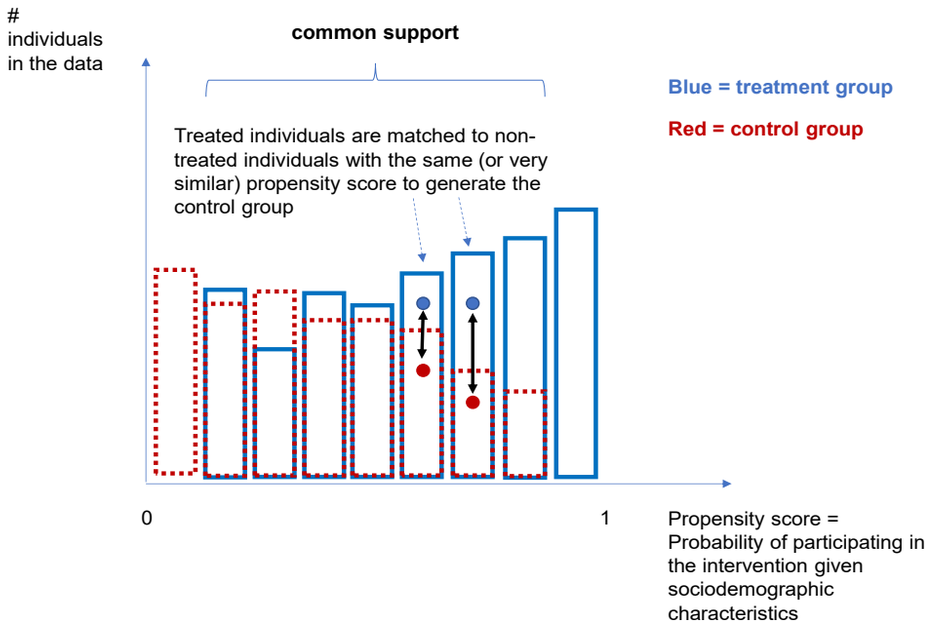
[79] In order to simplify this discussion, it is assumed that policymakers wish to know the effect of the treatment on those who actually received services from the programme (this is in many cases a subset of the target group that was offered the opportunity). This is called a 'treatment on the treated' (TOT) analysis.

[80] There is a wide range of potential approaches to matching on propensity score. For an accessible overview see Caliendo M., Kopeinig S., 2008.

[81] The extent of the region of common support has implications for sample size and the usefulness of results to policy, particularly where a large number of treated cases fall outside the region of common support.

propensity score matching can become a highly complex process with a range of issues to consider. Many of these issues are ignored here in order to ensure that the key principles are clear. A practical example, where an ESF evaluation used a matching approach, is presented in the next box.

Figure 8 Illustration of the propensity score approach



Selection based on observable data

The plausibility of the propensity score approach rests on the assumption, among others, that selection into treatment can be fully characterised by the observable data. In other words, that there are no unobserved differences between treatment and control groups that are related to results and/or the decision to participate in the intervention. The plausibility of this assumption is enhanced by incorporation of a rich range of variables into the estimation of propensity scores, the selection of variables being based on prior knowledge and theory. Specifically, in the context of labour market interventions the inclusion of individual past labour market histories is highly recommended when checking for potential unobserved differences.[82]

Box 15 An example of an evaluation adopting a matching approach[83]

The Impacts of ESF interventions financed in 2014-2020 for long-term unemployed in the Marche region [84]

A matching approach was used to evaluate the impact of a number of ESF interventions (internships, job fellowships, work experience in municipalities and training vouchers) on the long-term unemployed (LTU) financed in the Marche region of Italy. The interventions were not specifically targeted at the LTUs but they represented the majority of participants in all the ESF interventions analysed.

The aim of the evaluation was to determine the impact of ESF measures on the probability of being employed at 6, 9, 12, 15 and 18 months after the start of the interventions. Several outcome variables were used: the probability of being employed at a certain time after the interventions, the probability of being employed in an open-ended contract and the number of days worked during a period after the interventions.

[82] See Caliendo, Mahlstedt and Mitnik, 2017 and Kluve, Lehmann, and Schmidt, 2008.
[83] Other examples of practical application of methods (PSM, RDD and Diff-in-Diff) can be found in European Commission, 2020.
[84] See Pompili M., Giorgetti I., 2020a

A propensity score approach was adopted (Nearest-Neighbour Matching) to identify differences between treated and non-treated groups, and the four ESF interventions mentioned above were analysed separately. For each intervention the treated group was composed of participants who started the intervention before the end of August 2019 (526 for internships, 1058 for job fellowships, 236 for work experience in municipalities and 241 for training vouchers); the control group was composed of unemployed registered with the PESs in the period 2016-2018 with a spell of unemployment of at least 12 months (77,255 records).

Several variables were considered to calculate the PSM taken from the socio-demographic characteristics of treated and not-treated, such as: sex, age, citizenship, level of education, place of residence, date of entering the intervention and employment history back to 36 months before the intervention.

To measure the outcome variables for treated and control groups (before and after the interventions) data from the COB, the archive of firms' mandatory notification of labour contracts sent to the PESs, were used.

Results from the study were mixed, with positive impacts reported for internships and job fellowships, while negative effects emerged for work experience in municipalities and no significant effects for training vouchers[85].

3.2.3. Difference-in-differences

Key features of difference-in-differences

- Difference-in-differences is an intuitive approach that compares the before/after difference in outcomes in the treatment group with those in the control group
- Since the change over time in the control group measures what would have happened to the treatment group in the absence of intervention (the counterfactual), any additional difference in before/after outcomes in the treatment group gives the causal effect of the intervention
- This is a straightforward method that is practicable in many cases

Before and after the treatment

Either separately or in conjunction with matching, baseline (or pre-treatment) measures of result variables can be used to conduct difference-in-differences (DID) estimations. Here, the difference in a result before and after treatment in a control group is subtracted from the same difference observed in a treated group to arrive at an estimate of an intervention's impact. Again, selection of a plausible control group is essential. Impacts calculated on the basis of difference-in-differences are usually observed within a regression framework, which also accounts for other observed differences between treatment and control groups. Moreover, this approach checks for unobserved differences between the two groups which are fixed over time as well as those which vary through time but which affect both control and treatment groups equally (for example factors affecting the whole economy). Owing to this ability to check for some aspects of unobserved differences between treatment and controls, a difference-in-differences approach represents in most cases an improvement over a cross-section matching strategy. Figure 4 provides a visual representation of the difference-in-differences approach.

The x-axis represents the passage of time and the y-axis a scale upon which results are recorded. Results in this case might be wages. Average wages for the treatment group in the pre-treatment period are YT1, whilst for the control group they are YC1. In the post-treatment period wages are YT2 and YC2 for the treatment and control groups, respectively. Thus, the solid upper line

[85] For training vouchers, as underlined in the evaluation report, findings can be considered only as preliminary, since the sample considered is too small in relation to the total number of participants.

represents the change in wages among the treatment group, whilst the solid lower line that among the control group.

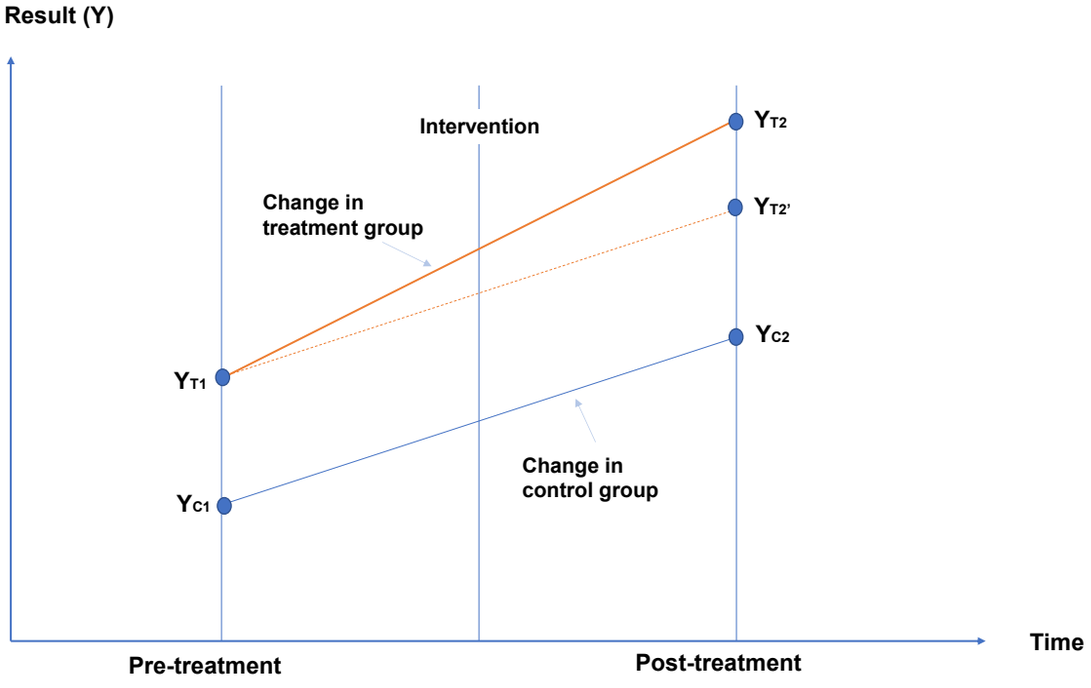
A crude estimate of the impact of the intervention would emerge from a comparison of wages in treatment and control groups in the post-treatment period, i.e., YT2 - YC2. This would, however, be incorrect as it would ignore differences in pre-treatment wages. One way to think about the difference-in-differences estimator involves viewing it as subtracting a pre-treatment estimate of bias from the post-treatment difference in results. As a result, the post-treatment difference in wages (YT2 - YC2) is adjusted by subtracting from it the pre-treatment difference in wages (YT1 - YC1) and therefore the difference-in-differences impact estimator can be expressed, very simply, as follows:

The logic of double difference

$$(YT2 - YC2) - (YT1 - YC1).$$

If the post-treatment differences in wages are not adjusted for pre-existing differences between treatment and control groups, biased estimates may result. Alternatively, as previously mentioned, the difference-in-differences approach can be thought of as subtracting the change in results among the control group from that change observed in the treatment group. The observed change in the control group is perceived as that which would have occurred in the treatment group in the absence of the intervention.

Figure 9 Illustration of difference-in-differences approach



The main assumption: common trends are equivalent in absence of treatment

In the simplest case, the main assumption upon which the difference-in-differences approach rests is that of common trends: i.e., time trends in results within treatment and control groups are equivalent in the absence of treatment. This assumption cannot be tested directly, although in cases where multiple pre-treatment measurements on the results variable are available for both treatment and control groups, and these measurements show parallel trends, it supports the plausibility of the approach. For an example of the difference-in-differences approach see the Box below.

Box 16 An example of an evaluation adopting a difference-in-differences approach

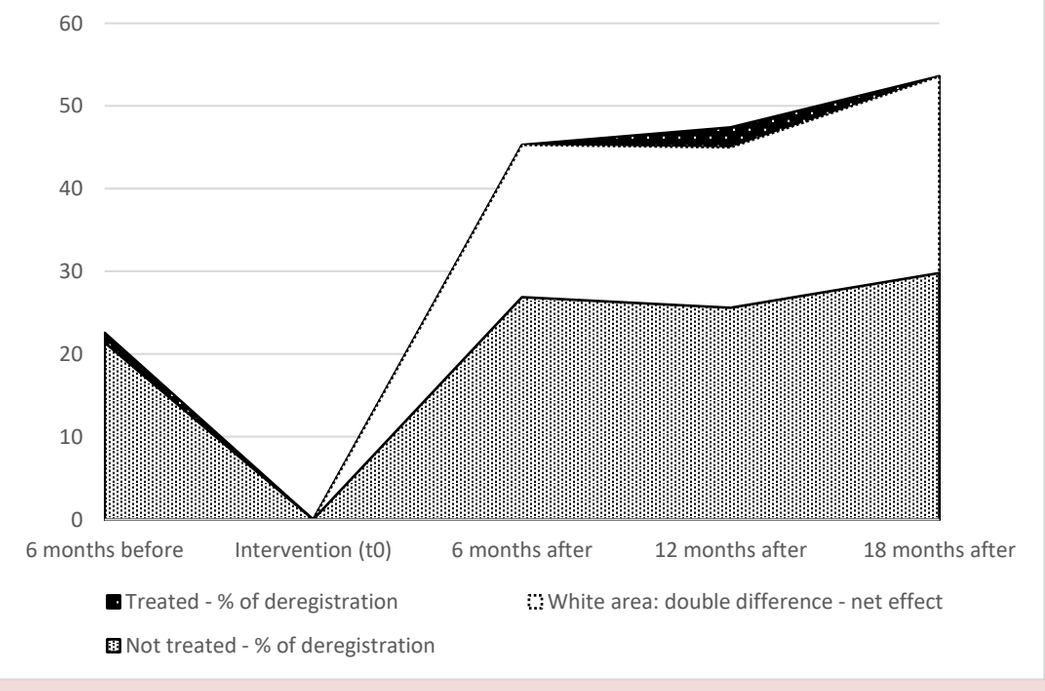
Evaluation of measures for older workers in Lubelskie financed by the ESF in the 2007-2013 programming period[86]

The main aim of the study was to evaluate ESF interventions for extending the economic activity of older persons in the context of unfavourable demographics in the Lubelskie region in Poland. The interventions were financed under the Human Capital Operational Programme 2007-2013.

The treated group, a sample of 145 individuals, consisted of people who were unemployed at the time of entering the interventions. The control group was taken from the anonymised registry data provided by Employment Offices, containing socio-demographic data and the history of registry events relating to the employment status of specific persons (data were obtained from eight Employment Offices and contained information on 67,102 persons). The outcome variable was the "status of being registered in the unemployment registry"[87] and this could be observed both before and after the intervention, allowing the implementation of a difference-in-differences approach. Employment status was measured at 5 points in time: 6 months prior to project participation, during project participation, and 6, 12 and 18 months after project completion.

The choice of control group in this difference-in-differences approach was enhanced, using matching methods to identify a non-treated person for each treated one, based on a number of observable covariates, such as gender, age, level of education, unemployment rate at the place of residence.

The analysis showed that support for the unemployed aged 50+ produced a positive net effect on employment: the treated 50+ unemployed had over a 2.5 times higher chance of deregistering as unemployed than a non-treated person. The figure below shows the evolution of the outcome variable for the treated group and the control group and the estimated net effect after considering the parallel assumption of the difference-in-differences approach.



[86] See Re-source Pracownia Badań i Doradztwa, 2015.
[87] Although to be unregistered as "unemployed" is not identical to being "employed", such an assumption was necessary in order to conduct the study

3.2.4. Regression discontinuity design

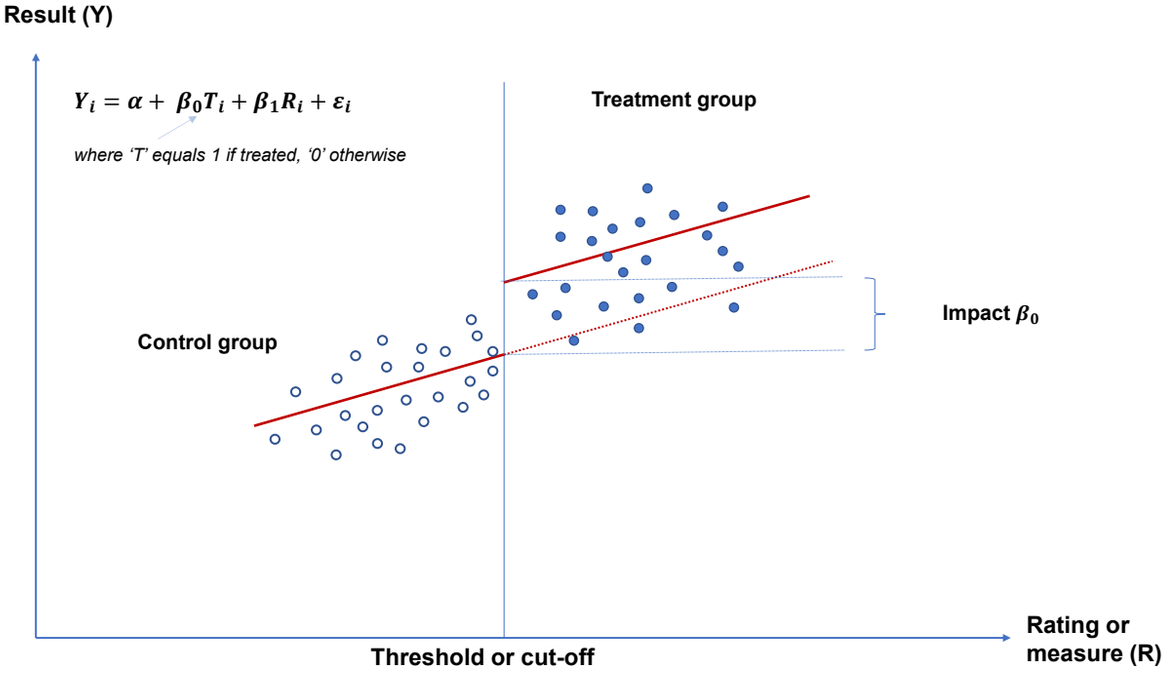
Key features of regression discontinuity

- A regression discontinuity design is based on the idea that a specific value ("cut-off") of a score or rating determines whether an individual participates in the intervention or not
- Individuals close to that value are then considered comparable, with the only difference that those on one side of the cut-off participate (the treatment group), and those on the other side do not (the control group)
- Comparing these groups typically gives a precise and intuitive measure of the intervention effect; but the method is only applicable if a selection process based on a score or rating is in place

A cut-off point divides treated and control individuals

A **regression discontinuity** approach may be adopted when access to an intervention is determined by a cut-off point along a continuous rating, scale or measure. For example, access to training might be determined by performance in an aptitude test, with those scoring above a specified threshold (or cut-off) receiving training, whilst those who score below the threshold do not. For the approach to be valid, the cut-off point should be determined without knowing the scores of potential trainees; candidates immediately around the cut-off point will be very similar to one another, but for the fact that those just above it will take part in the intervention while those just below will not. Results for those above and below the cut-off can be compared to obtain an estimate of the intervention's impact at the cut-off point.

Figure 10 Illustration of the regression discontinuity approach



Sharp or fuzzy discontinuity

A regression discontinuity design (RDD) can be implemented where the cut-off point either identifies the treatment group completely (with full compliance), in which case a *sharp* discontinuity is obtained, or where, under certain conditions, not all those on a given side of the cut-off point comply strictly with their assignment to treatment (a *fuzzy* discontinuity).

Figure 10 above presents a stylised example of a regression discontinuity design. This is the simplest illustration of a sharp discontinuity. The intervention produces constant effects at each value of the rating and impacts are estimated using a linear regression model (there are no issues regarding the functional form of the impact regression). In practice, the analysis will invariably need to be significantly more sophisticated than that presented in Figure 10.

The dots in Figure 10 represent individual units, for example trainees. The x-axis records the rating or measure used to allocate trainees to slots on the training course. Individuals scoring to the right of the solid vertical line indicating the threshold on this rating or measure (an aptitude test for example) enter training and form the 'treatment group'. Potential trainees scoring below the threshold on the rating or measure do not enter training and form the control group.

The key point is that the rating used to allocate the target group to either treatment or control conditions is a continuous quantitative variable measured prior to treatment and an individual is admitted to the training scheme based on whether their score is above or below a pre-defined cut-off or threshold.

The result is plotted on the y-axis. Essentially, the treatment impact is identified through estimating a linear regression model on the data (given the assumptions above); i.e., regressing the result variable against the rating measure along with a dummy variable (a treatment indicator) which defines whether a score is below or above the cut-off point (i.e., whether the individual is assigned to the treatment or control group).

Such an impact regression equation is illustrated in Figure 10. The effect or impact of training in our example is obtained from the coefficient on the treatment indicator, i.e., β_0.[88] This effectively shows whether there is a break or discontinuity around the cut-off point, indicated in Figure 10 by a shift upwards in the regression line at the threshold or cut-off. In this example, a positive impact of training on the result is observed.

An alternative way of understanding the impact estimate is to consider the dotted line extension to the control group line depicted in Figure 10. This can be thought of as a counterfactual estimate for the treatment group - the relationship between the rating and result measure which would have prevailed in the absence of the intervention - the difference between this dotted line and the trend line for the treatment group represents the treatment effect or impact. Notice that in the absence of treatment there is no discontinuity in the line and we assume that in such a case the result varies continuously with the rating or measure.

Good internal validity but generalisation may be limited

The regression discontinuity approach works because observations in treatment and control groups close to the cut-off point are similar to each other but for the fact that those above the cut-off point, in this example, receive training, whilst those below do not. The situation is therefore not unlike randomisation for observations close to the cut-off point. There is, however, one considerable limitation. In most applications, impacts estimated using an RDD approach can only tell the policy maker about effects for individuals close to the cut-off or threshold. The degree to which generalisations can be made about those further away from the threshold may be limited.

[88] In a simple case this would be the effect of intention to treat at the cut-off point (see Bloom H. S., 2009)

RDD can be a useful approach where individuals are allocated to an intervention on the basis of need measured on a continuous rating or score. However, analysis can become complex where the cut-off point is fuzzy and there is non-compliance, and where issues of functional form in the impact regression model exist. Effectively, a range of assumptions need to be invoked and the validity of these assumptions cannot always be confirmed.

The next box shows a practical example, where a regression discontinuity approach was used in an evaluation of a Youth Guarantee scheme.

Box 17 An example of an evaluation adopting a regression discontinuity approach

Vocational Training for Unemployed Youth in Latvia: Evidence from a Regression discontinuity design[89]

Researchers from the Joint Research Centre of the European Commission used a regression discontinuity approach to assess the impact of a vocational training (VT) programme on labour market outcomes of unemployed youth in Latvia, funded under the Youth Guarantee scheme for the period 2014-2020 and targeted at young people aged between 15 and 29 not in education, employment or training (NEETs).

The data used was drawn from the Latvian State Employment Agency (SEA), which provided information on both participants and non-participants registered as unemployed on specific dates, and was matched with data from the State Revenue Service (SRS), which contains information on the income of individuals on specific dates before and after the programme (January 2012 and June 2017). After cleaning the datasets, the final sample of the treated group added up to 898 individuals and that of the control group to 10,717.

The evaluators relied on a Latvian government rule that gives participation priority to unemployed people under 25 in the VT programme. Therefore, age is the running variable which determines the probability of participation in the VT programme, the age of 25 representing the threshold below which participation in the programme drastically increases due to the priority rule. Since individuals have no control over their age, the allocation to the VT programme can be considered random around the threshold. The priority rule and the fact that participation was voluntary required the researchers to apply a Fuzzy Regression Discontinuity Design (FRDD). The identification strategy relied on the fact that people close to the cut-off point (age 25) were similar to each other but for the fact of participation in the VT programme.

The findings showed that the impact of probable future employment and a monthly income up to 3.5 years after entering the programme is positive but statistically not significant. However, a positive effect of the priority rule on programme participation was observed. Since the validity of the FRDD relies on the fact that potential participants cannot control the running variable (age), several tests were implemented to confirm this effect.

[89] See Bratti M. et al., 2018.

3.2.5. Instrumental variables

Key features of instrumental variables

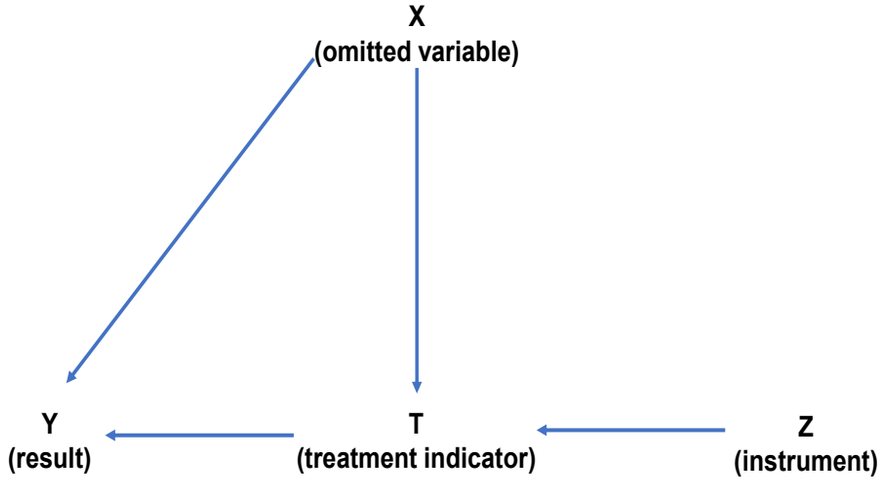
- The idea of Instrumental variables is that some (pre-determined / exogenous) process determines participation, but it is not the actual selection procedure
- The pre-determined process, and its key variables, can be used to take into account any self-selection in the decision of individuals to participate or not.

How exogenous factors affects participation

For the **instrumental variables** (IV) approach, selection into treatment should be at least partially determined by an exogenous factor which is unrelated to results other than through the treatment. As such, the exogenous factor has influence on participation, but not directly on the results. Typically, such exogenous factors can be administrative errors or oversights, or other random variations in the treatment received.

Figure 11 below illustrates the instrumental variables approach. Four variables are shown in a highly simplified causal system. The variables represent data collected from a population hypothetically targeted by a training intervention (both those who receive training and those who act as controls).

Figure 11 Illustration of an instrumental variables approach



Y' represents the result under consideration. In the case of a training intervention this could be earnings. 'T' is an indicator which reveals whether an individual has taken up training.[90]

'X' is an omitted variable which is not observed but which is related to both 'Y' (the result) and 'T' (the treatment indicator), extending the idea of a training programme, a baseline measure of ability for example. In this case, ability is related to both participation in training and earnings. For example, more able

[90] In other words, there is full compliance, and all those in the treatment group participate in training

members of the target group may choose to take up training as well as enjoy higher wages.

The existence of 'X' motivates the search for an instrument and means that the impact of training on earnings is confounded by its existence. In other words, the estimate is biased because of the existence of X and the fact that it is unobserved and cannot be directly accounted for in the analysis.

The identification of a valid instrument is fundamental

Finally, the variable 'Z' is an instrument which, according to Morgan and Winship[91], can be thought of as a shock to 'T' which is independent of 'X'. For this reason, there is no line in Figure 11 which links Z with X. Moreover, Z only affects Y through T, there is no other pathway through which Z affects Y. This means that Z can be used to generate variation in T (the treatment) that is uncorrelated with the confounding variable X. As a result, an unbiased measure of the effect of T on Y can be obtained through exploiting this variation.[92]

The simplest circumstances in which an IV approach might be taken are described below, leaving aside here many of the complexities involved. In practice, it is often difficult to find a convincing instrument. The plausibility of different potential instruments is highly context-bound and the underlying identifying assumptions can generally not be tested statistically (actually, the required correlation between Z and T can be tested statistically, whereas the "independence" of Z and X cannot). For example, one strategy might be to use the distance from centres where training is provided (the physical location of the training course) to a trainee's home as an instrument in estimating the effect of training on trainees' net earnings. It might be observed that trainees that live closer to training centres are more likely to participate in a training intervention and that the distance between a trainee's home and a training centre is unrelated to other determinants of net earnings and participation in training (for example human capital measures). The only pathway, therefore, through which this distance measure might affect net earnings, is through its effect on training.[93]

Instrumental variables can be used in a wide range of contexts. Estimates can be arrived at through a variety of estimation approaches, depending on the response variable. So far, this approach has not often been used in ESF evaluations. In Box 18 below, we illustrate an example of the analysis of causal effects of a policy applying an IV approach.

[91] Morgan S.L., Winship C., 2014.
[92] The causal effect of T on Y is calculated in the presence of an instrument through estimating the relationship between Z and Y, and dividing this by the estimated relationship between Z and T
[93] Interpretation of findings from such an analysis may be complicated by whether the instrument is correlated with variation in treatment effect (see Bryson,et al, 2002)

Box 18 An example of a study adopting an instrumental variables approach

Employment effects of language training for unemployed immigrants[94]

Acquiring competence in the host country's language is an important factor for immigrants in achieving a high level of integration into the labour market.

In a recent paper (Lang J., 2021) the author uses an instrumental variable approach to estimate the causal effect of a German language training programme for professional purposes on the employment performance of immigrants participating in the courses two years after enrolment. The programme was implemented by the German Federal Office for Migration and Refugees (BAMF) and financed by the ESF; the ESF-BAMF programme was operative until the end of 2017. The courses had several components: German language training, professional skills building and work placements. People with a migration background and active in the German labour market were eligible, irrespective of nationality and date of immigration.

To address the unobserved heterogeneity in terms of language skills, the author uses the exogenous variation in local language training intensity at the job centre level. This is considered a valid instrument since the job centres have broad discretion in the implementation of different mixes of programmes depending on their "policy styles", and this variable is exogenous to jobseekers' employment outcomes. Courses implemented in 2014 were analysed.

The evaluation relies on a rich set of administrative data, the Integrated Employment Biographies (IEB), a merged database of administrative information of the German Federal Labour Agency. The IEB includes information on participation in language training, detailed information on employment history (except for self-employment), job searches, transfer payments made during the unemployment period, and individuals' personal data. Other data (WGH, Werdegangshistorie) were also used so as to have detailed information on participation in the language courses provided by BAMF, and to cover data missing from IEB where possible with information on self-employment and parental leave episodes.

Findings show that after a period of lock-in effects lasting just over a year in general, two years after the intervention, immigrants participating in the training programme have a 9 percentage points higher probability of being employed than the control group, and this probability is not restricted to unskilled jobs. It was also found that early provision of language training (soon after arrival in Germany) is beneficial for the integration of immigrants in the labour market.

[94] See Lang, J., 2021.

Table 5 Comparison of key features of main CIE approaches

Approach	Key features	Advantages	Data requirements	Limitations
Randomisation - experimental approach	- Individuals eligible for participation are randomly allocated to a treatment and a control group - Randomisation ensures that both groups are identical (in general) in all relevant characteristics - Hence, the control group answers the counterfactual question and the difference in outcomes between treatment and control gives the causal effect of the intervention	- If implemented correctly, estimates of impact are 'unbiased' - Results are transparent and easily understood - Findings are less subject to qualification and doubt - Variety of design variants available to cope with a range of policy contexts and intervention circumstances	- Basic requirement to control selection into the intervention via randomisation - A record of who has been allocated to which group - Collecting baseline data is essential - Result measures need to be recorded for both treatment and control groups	- Often treated and control group do not comply with the allocation rules to the treatment - Participants' consent is often required - Randomisation can influence the selection of those who participate in/apply to an intervention - Being aware of their assignment status can alter participants' behaviour and influence results - Ethical concerns - Considerable planning and design requirements - Can be costly (though not necessarily so)
Matching (propensity score)	- Matching mimics randomisation by constructing ex-post a control group that is as similar as possible (in general) to the treatment group in all relevant characteristics - Different from randomisation only observable characteristics (age, gender, education, etc.) can be matched, while unobservable characteristics (e.g., motivation) cannot be take into account - The validity of the approach thus crucially depends on data availability	- Requires good knowledge of selection processes, but does not require direct control over selection into the intervention - Can be applied retrospectively, if the right data are available and in a variety of contexts - Technically a semi-parametric method of estimation; requires fewer parametric assumptions (for example, no need for standard regression assumption). - Can be used to estimate multiple treatment effects	- Accurate identification of intervention participants - Data sources from which to make sample - Clear concept of participation and good understanding of selection into treatment - Rich data, ideally collected at baseline from which to select the match - Result measures of the intervention for participants and controls	- Requires considerable amounts of data that allow a full characterisation of the selection process - Validity depends on quality of controls and their careful selection and the degree of common support - Relies on the assumption that selection into the intervention can be characterised adequately by observable data - The range of different available approaches to matching requires sensitive analysis - Results can be complex to explain and interpret, and potentially ambiguous
Difference-in-	- DID is an intuitive approach that compares the before/after difference of outcomes in the	- Checks for some aspects of unobserved differences between	- Data requirements are similar to other approaches but with the additional	- Requires assumption of common trends in results between participants

Approach	Key features	Advantages	Data requirements	Limitations
differences (DID)	treatment group with the before/after difference in the control group - Since the change over time in the control group measures what would have happened to the treatment group in the absence of the intervention (the counterfactual), any additional difference in before/after outcomes in the treatment group gives the causal effect of the intervention - This is a straightforward method that is applicable in many cases in practice	participants and controls - Can be used in conjunction with matching - Works with data from pre- and post-intervention, such as panel data (individual data over time) or repeated cross-sectional data (data on individuals collected at different points in time)	requirement for pre- intervention measures on results - In order to test main assumptions multiple pre-treatment observations on results are required for both treatment and control groups	and controls to be invoked - Analysis can become quite complex and open to misinterpretation - Rich pre-treatment data on results required to test assumption of common trends - Cannot be used to estimate multiple treatment effects[95]
Regression discontinuity design (RDD)	- RDD is based on the idea that a specific value ("cut-off") of a score or rating determines whether an individual participates in the intervention or not - Individuals close to that value are then considered comparable, with the only difference that those on one side of the cut-off participate (the treatment group), and those on the other side do not (the control group). - Comparing these groups generally gives a precise and intuitive measure of the intervention effect; but the method is only applicable if a selection process based on a score or rating is in place	- Both sharp and fuzzy approaches to RDD are available. - Can provide unbiased impacts of treatment effects subject to certain conditions	- The choice of cut-off point needs to be independent of the values on the rating given to each member of a target group - Data is required on individuals for both treatment and control groups in terms of the rating or measure, the threshold or cut-off and results	- This approach is not valid without a continuous measure or rating which determines treatment - Analyses can become complex and uncertain where issues of the functional form of the impact regression become prominent, where there is non-compliance or where the size of the sample around the cut-off is limited - There can be dangers in interpreting findings and extrapolating generalisations.
Instrumental variables (IV)	- The idea of IV is that some (pre-determined / exogenous) process exists that determines participation, but is not the actual selection process - Then that pre-determined process can be used to take into account any self-selection in the actual decision of individuals to participate or not	- Can provide high quality estimates of, or evidence for the existence of, causal effects - Solves the problem of omitted variable bias (or selection bias) - Can be applied retrospectively	- Requires baseline data, data on results and participation in the intervention, and in addition, that an instrument can be identified - An instrument needs to be related to participation in the intervention and only affects results on this basis. The instrument should not be correlated with any other determinants of results	- Can be difficult to find a plausible instrument - Can be difficult to explain to non-experts - Interpretation of results not straightforward; limited testability of identifying assumptions

[95] See Frolich, M., 2004.

Chapter 4. Moving the CIE Agenda forward

This Guide seeks to encourage and support MAs in conducting more widespread and good quality CIEs. It provides guidance to those who are responsible for planning and commissioning impact evaluations of ESF+ co-financed interventions. Thus far, the focus has been on planning and implementing a CIE, and a number of key questions that require consideration have been discussed. However, the 2014-2020 programming period highlighted a number of other 'wider issues' as well as challenges that have to be dealt with to strengthen ESF+ evaluations and the use of CIEs.

Wider issues to tackle

This section of the Guide makes some suggestions for tackling these 'wider' issues. In particular, steps to address the following are discussed:

- Lack of knowledge of CIE approaches within MAs and among the wider MS policy-making community;
- A lack of external, suitably qualified and experienced contractors in MSs, able to undertake CIEs;
- Addressing legal barriers that need to be confronted generically across CIEs;
- Moving forward towards greater planning of CIEs;
- Broadening the scope of CIEs

4.1. Improving levels of understanding among stakeholders

For the programming period 2021 - 2027, the CPR stipulates in art.44(1) that" *(the MA) shall carry out evaluations of the programmes related to one or more of the following criteria: effectiveness, efficiency, relevance, coherence and Union added value, with the aim to improve the quality of the design and implementation of programmes. Evaluations may also cover other relevant criteria, such as inclusiveness, non-discrimination and visibility, and may cover more than one programme".* As discussed in the previous sections, this means that appropriate evaluation capacity must be available in both the MA contracting out CIE and the evaluators applying for the contract.

In some cases, the Evaluation Helpdesk services recorded, a lack of capacity for carrying out CIEs – that is, tendering and accompanying the implementation – in some administrations. This made it difficult for evaluators to conduct CIEs because clear evaluation questions, basic data availability and well-informed implementation planning had not been identified in advance.

Stimulate demand and supply of CIE

Sometimes, especially in small countries, capacity to conduct a CIE is lacking in the consultancy market as is a supply of technical expertise. Consequently, there is a widespread need to stimulate both demand for and supply of CIEs. The supply issue may improve as MAs and MSs start to commission CIEs, or make their requests for such studies known. The speed of response to increased demand for CIEs will depend on pre-existing skills, experience and the existence of institutions within the MSs capable of implementing such

approaches. However, stimulating demand can also be partly achieved by improving the knowledge and understanding of CIE methods among those working in MAs.

Developing training in CIE methods

One solution to this problem is for MAs to run training courses in CIE methods for their staff. Training should focus on the benefits to MAs of adopting CIE methods. Moreover, issues of accountability and learning what works should be stressed. A suggested course outline is provided in Annex 2.

4.2. Capacity development

One other issue raised during the 2014-2020 period, and mentioned in the sections above, was the need to develop capacity to carry out CIEs within MS research/academic/consultant communities. In some cases, it was apparent that the skills required to do so were available within MSs, but that those with the skills had faced barriers to applying them within an evaluation context (e.g., limited access to useable data, or problems in identifying a reasonable control group).

Strengthening institutions and building communities of practice

There are a number of steps that can be taken to develop supply of evaluation services. Many of the issues raised apply equally to CIEs as to evaluations more in general. Three steps are commonly taken to improve evaluation supply:

- Build relationships with academic institutions, in particular universities;
- Develop and strengthen an independent community of consultants;
- Support the development of a professional evaluation community.

Universities

Developing academic skills

Developing links with universities is important for two reasons. First, academic staff at universities may possess the skills and knowledge required to conduct CIEs. For example, many micro-economists, econometricians, quantitative sociologists or psychologists have the types of skills necessary. In many MSs the required skills may be available but those who have them have not previously thought to apply them to the evaluation of interventions. There may be a lack of incentive for them to do so that will need to be addressed.

In some MSs there is a tradition of academic researchers actively engaging in applied policy research. In this setting, academics will be familiar with working with government and MAs. In other MSs, where universities and academics are not as engaged in applied work, a culture change may be required. One successful method of developing a supplier base within the university sector is for MS authorities and MAs to core-fund the costs of research centres dedicated to CIE methods.

Training the next generation

Universities and academics can also play a role in training the next generation of evaluators. When working closely with universities, it may be possible to encourage them to include programme evaluation methods within their curricula, and as part of this development, ensure CIE methods are covered within teaching programmes. In some MSs, universities may also play a role in running continued professional development courses on impact evaluation and CIE methods. This can be aimed at policymakers, technical specialists within MAs, as well as other potential suppliers such as independent consultants. MSs might consider providing funding for such training.

Independent consultants

Developing the market

For some large-scale evaluations there is an international market. This is certainly the case for large CIEs. However, many MSs will want to develop domestic capacity to conduct CIEs. One strategy toward achieving this can be through establishing strategic alliances between potential domestic suppliers and international consultancies.

Several suggestions for developing a domestic supplier base to undertake CIEs are set out below, and may be applied by MAs (or other bodies) commissioning CIEs:

- Insisting that consortia or partnership bids always include some local consultants;
- Scaling evaluation contracts in ways that relatively small, low-risk evaluations can be undertaken by new, national entrants to the evaluation market;
- Ensuring that technical and financial requirements associated with bidding for evaluations are not too restrictive and allow the participation of new entrants;
- Emphasising technical and know-how criteria in the selection rather than complex administrative procedures with which less experienced consultants may not be familiar;
- Holding briefing meetings with potential consultants to answer questions and encourage bids in a competitive environment;
- Support for networking among relatively isolated evaluation consultants, so as to encourage team-building, consortia formation and other professional networks and associations, also at international level;
- Acknowledgement by evaluation commissioners that they may need to take a more hands-on management approach to new contractors to speed up their acquisition of knowledge and experience.

Box 19 An example of a project aimed at strengthening CIE culture and capacity

In the 2014-2020 programming period an ESF project was carried out in Sweden with the aim of improving culture and evaluation capacity of the Public Employment Services (PESs). The project, called" Evidence based EU funded projects", ran from 2016 to 2018 and involved PESs managers and other staff in workshops and other training actions in the field of evaluation.

The focus of the activities was not on technical issues but on the importance of an evaluation, the main steps of an impact evaluation, the importance of good data and the quality assurance of an evaluation. A website was set up and educational material uploaded.

The main results of the project were the improvement in the knowledge of PESs managers in the field of evaluations, their openness to the practice of evaluation and their cooperative attitude. In concrete terms, according to the interviewed person "the effects of the projects on the number of impact evaluations carried out" are not yet high, but some progress is evident, considering that in the previous programming period 2007-2013, no CIEs were carried out in Sweden.

The evaluation of the 'Ung framtid' (Young future) project, which adopted a randomised approach, is a concrete example in practice, which involved PESs managers in the evaluation process[96].

[96] Information is based on an interview with a government official from the Swedish PES Arbetsförmedlingen.

Professional community

Developing professional communities

It is important to develop professional evaluation communities within MSs, with space for the discussion of CIE methods and sharing experiences. The development of professional communities is important for mutual support and learning, and also for the maintenance of quality standards. A useful strategy could be to develop links with the relevant national evaluation societies and encourage them to promote CIEs of ESF+ interventions in training events, specific conferences or seminars, or awareness-raising sessions.

Sharing experience

Utilising existing forums

The EC is keen for more rigorous ESF+ impact evaluations to be conducted, and CIE has been widely recommended in the 2014-2020 period[97].At present, it is possible to affirm that the use of CIEs has increased, but they are still limited in a number of MAs and MSs. Sharing experiences in CIE methods is one of the foremost means to develop capacities and support the diffusion of CIEs throughout the EU. Existing forums of mutual learning such as peer reviews of employment and social inclusion policies, and communities of practice within ESF+ should be utilised for this purpose. Initiatives by individual MAs or MSs, such as international conferences or seminars, could also favour the exchange of experiences.

4.3. Confronting legal barriers

Removing legal barriers to data access

One of the most significant and substantial problems encountered by researchers conducting CIEs across MSs is gaining access to data. In particular, researchers often encounter legal barriers that aim to protect the confidentiality of persons represented in data sets. As shown above regarding the features of the GDPR, the answer to addressing these issues lies in undertaking wider reforms, and concluding agreements that enable relevant data to be made available to evaluators in a controlled manner, on an ongoing basis.

Creating analytical datasets

Analytical versions of administrative datasets could be constructed on a regular basis from administrative data that are held by MS authorities, documented and deposited in an archive with controlled access. Approved contractors can extract data from such holdings with authorisation. Should it not be possible to obtain specific consent, data would be fully anonymised with encrypted personal identifiers to respect GDPR regulations. Data holdings like this have been created in several countries. However, where access is still an obstacle due to different interpretations of privacy rules, a national initiative at government level should promote agreements and systems able to deliver data for research purposes in a relative short period of time.

Creating data labs

If concerns over confidentiality of personal data persist, consideration should be given to the establishment of data labs. Here evaluators working on administrative datasets would be given access to records only at secure locations, where access to data is strictly monitored and controlled. Data

[97] The CPR Regulation for the 2014 – 2020 period (Regulation 1303/2013) – annex XI asked for an “effective system of result indicators necessary to monitor progress towards results and to undertake impact evaluations”. This requirement was part of the ex-ante conditionalities for the 2014-2020 programming period.

would have to be processed and analysed on site, and only the results of any analyses could leave the premises.

4.4. Moving towards more prospective approaches

A common feature of the small number to date of CIEs conducted of ESF-financed interventions is that they have been retrospective in nature rather than prospective. What this means is that expert evaluators have been commissioned to conduct evaluations of interventions that have been developed without any consideration of evaluation, and in some circumstances where little or no planning for an impact evaluation has taken place. This means that evaluators have had to construct data sources in time-consuming, expensive and sub-optimal ways, responding to the data that happened to be available, rather than data sources constructed with impact evaluation in mind.

Prospective approaches include evaluators from the beginning

In contrast, a prospective approach would involve evaluators in planning for a CIE at the earliest opportunity and would enable interventions (either new or existing) to be influenced, often in quite subtle ways, making them more amenable to CIE. Planning in advance for a CIE can mean the difference between being able to conduct a rigorous evaluation or not being able to do one at all. Involving appropriately trained internal staff or engaging external expert contractors early in the life of an intervention or when funding decisions are being made means that:

- Appropriate recordkeeping can be integrated into the delivery of programmes and interventions;
- Requisite data sources can be identified early and access and data protection issues dealt with in good time;
- Baseline data collection can be specified and surveys administered if required;
- Practical issues relating to how participants are recruited into interventions can be addressed in ways which mean that recruitment processes are more consistent with rigorous evaluation.

The involvement of evaluators trained in CIE methods (be they internal MA evaluators or externally commissioned experts) in the process of developing new ESF+ interventions or in decisions concerning existing interventions, would enable planning for impact evaluation to commence at the beginning of the programme period, as well as reaping significant benefits in evidence-based policy decision-making.

4.5. Broadening the scope of CIE

A final, very important issue concerns the policy coverage of CIEs and involves different dimensions: the policies covered, the analysed outcomes and the completeness of the interpretation of findings.

More focus on CIEs in social inclusion and education policies

In the 2014-2020 period the diffusion of CIEs was focused almost exclusively on active labour market policy and effects relating to employment status. This can probably be explained by the fact that administrative data on employment are generally easier to access, and these data provide the key outcome variable for the analysis of many active labour market policies.

Education and social inclusion policies have not, or only very rarely, been subjected to a CIE despite their importance in ESF strategy and funds allocation; some of the few available examples from Spain, Poland and Portugal are reported in box 20 below.

Involving additional actors and developing appropriate data sources

To counteract this shortage of CIEs and improve knowledge of these important policies, a range of combined initiatives are necessary:

- Planning and preparing CIEs on social inclusions or education measures in advance so as to assess data availability in time; activate the necessary collaborations between different actors to increase that availability; and identify treatment and control groups at an early stage.
- Promoting the involvement of the data owners in CIE design; a broader involvement in CIEs of the administrations responsible for education and social policy, where different from the MA, might produce a stronger and more generalized commitment to evaluate and make data available.
- Stimulating institutional agreements and software tools to make administrative data on social conditions and education usable for CIEs. To this aim the involvement of the national statistical office, as a 'bridging' institution and facilitator, can be important, as these offices have the required skills and already elaborate many administrative datasets to produce national statistics;

Supporting relevant data collection from participants in treatment and control groups at the beginning of the intervention so as to have a 'before and after' consistent dataset. Where administrative data are not available or pertinent, a specific survey can provide the necessary data but must start at the beginning of the intervention and also involve the control group to ensure the necessary comparisons.

Box 20 Examples of evaluations in the field of education

ESF interventions against early school leaving implemented in Asturia[98]

Under a general evaluation of the OP Principado de Asturias ESF 2014-2020 a specific impact analysis of the Diversificación Curricular y de Mejora del Aprendizaje y del Rendimiento (PMAR) was carried out. PMAR is an ESF measure implemented under IP10.i, aimed at combatting early school leaving. Under PMAR students are divided into specific groups (from 8 to 15 students, but of different sizes if there are special circumstances), to study in the fields of Linguistics, Social Sciences, Science and Mathematics, as well as Foreign Languages (the remaining subjects are taught in the main class). The measure was implemented in the academic year 2016-2017, in the second and third year of secondary school (ESO). Students who have repeated at least one course at any stage and those who, having completed 1st ESO, but are not in a position to pass to the next year, are all eligible to participate. The final selection is made upon the individual assessment of students (academic and psycho pedagogical) carried out by the teaching teams.

A total of 1,053 students participated, 512 from the second year of ESO and 541 from the third. These students were compared with others with learning difficulties[99] in order to select controls comparable in terms of eligible criteria to the treatment students (total 3,852 individuals).

A matching approach was applied, using the administrative data from the SAUCE database, (Sistema Informático para la Administración Unificada de Centros Educativos) provided in an anonymised form by the Ministry of Education. These data were used to calculate the propensity score[100] and to measure the

[98] See Diaz J.M. et al., 2019

[99] These were identified as students who in the academic year 2015-2016 were in their first and second year of secondary school, who completed the course with one or more failed subjects and who had repeated at least one course in their school career.

[100] Variables to match the two groups were the following: academic performance during the previous academic year, type of school

outcome variable, which is the successful passage to the next academic year. The evaluation found a positive and statistically significant effect of the interventions: the second-year treatment students show a success rate 18 percentage points higher than the control group, for third-year treated students the difference is about 16 percentage points. In both groups the impact is more positive for females than for males.

ESF interventions to improve vocational education in Podlaskie[101]

The study examined the effects of the project "Good Profession - Great Life", which aimed to promote vocational education and training in the Podlaskie region in Poland. The project was targeted at staff of lower secondary schools, students and parents, and a range of activities was implemented both for the institutions and people (tailored marketing and communication strategies for schools, training for teachers, cooperation with companies and local industry leaders, mentoring for students and parents, vocational and educational counselling for students). The project was carried out between 2017 and 2019.

The treated population consisted of students who participated in the evaluated project. Out of 9403 students, 1500 were selected by stratified random sampling, and 200 students participated in the study's survey. Information on students who participated in the project was sourced from the monitoring data. The control group was specified as students with similar characteristics to those in the treated group, but who did not participate in the project. A control group of roughly twice the number of the treatment group was planned and selected at random from the national population registry (PESEL – a registry which gives each Polish citizen an individual number). However, due to difficulties caused by Covid, the planned methodology was not feasible. Eventually, the snowball (chain-referral) technique was used as a primary sampling method, resulting in 401 conducted interviews, out of which 384 were used in the study.

Data on outcome variables[102] for both the treatment and control groups were collected via questionnaire surveys, using personal, telephone or web interviewing. The study used propensity score weighting as a counterfactual method to assess impact (only at a general level, not for specific sub-groups). The counterfactual analyses conducted in the study did not provide conclusive findings. The statistical significance of the results was unsatisfactory.

Evaluation of the higher education grant system for less privileged students in Portugal[103]

A second evaluation, not yet officially published, examines a grant programme for students from low-income households aimed at favouring access to higher education and increasing the level of attendance. The grant is supported by the ESF (since the 2007-2013 period) in the regions of North, Centre and Alentejo and by the State in the other regions. Eligible students are those with resources below a specific threshold (between 6,8 and 7,9 thousand euros per capita income) and a minimum number of credits obtained in the previous year of study for students not in their first academic year. The same eligibility rules apply whether enrolling in public or private universities. The amount of grant is proportional to the household income.

Since 2011 the programme has involved about 70,000 students each year, but the evaluation covers the period from 2012 on, given availability of data from that year. Furthermore, the evaluation focuses on students applying for the grant for the first time and enrolled in their first year of a degree course (Bachelor or Master); eligibility is determined only by the income criterion[104].

Two administrative datasets were used: data on students applying between 2012 and 2018 (provided by the Directorate-General of Higher Education in Portugal (DGES)), which were merged through a unique identifier with another dataset containing information on academic careers and progression (provided by the Directorate-General for Statistics and Science (DGEEC)). The final sample under analysis is composed of 156,002 students, out of which 130,602 were treated and 25,400 not The income threshold (the running variable) allowed the evaluator to apply a regression discontinuity design and several outcome variables were measured: in the short term, the percentage of students still enrolled at the end of the first year, and

(public vs. private), enrolment in other additional course during the previous academic year (yes or no), sex, average income of the municipality of residence, place residence, country of birth (Spain vs. rest of the world).

[101] See Małgorzata Z. et al., 2020.

[102] Outcome variables (based on survey responses) were the following: attendance at a technical school at the time of the study, attendance at a 1st degree trade school at the time of the study, attendance at either a technical or 1st degree trade school at the time of the study, applying to a high school, applying to a technical school, applying to a 1st degree trade school, applying to a school providing vocational training (either a technical or 1st degree trade school), the school meeting a student's expectations to a very large extent, a vocation being definitely in line with a student's interests.

[103] See Guthmuller S., Meroni E.C., not yet published.

[104] The analysis of the full sample of students (not only those of the first year) is envisaged, though not completed yet.

credits obtained at the end of the first year; in the longer run, the probability of graduating, the final mark obtained and the number of years to graduation.

In the initial period, the grant shows positive effects in terms of enrolment rates but not in terms of credits obtained; in the longer period, the treated students graduate more, in less time and with higher marks than the control group. The effects were higher for males, students residing in territories covered by the ESF and for students enrolled in Masters and in state universities.

Increasing the number of meta-evaluations

The usefulness, as well as the explanatory capacity, of the meta-evaluations has been underlined before. A general increase in CIEs and, consequently, a possible increase in meta-evaluations would strengthen the debate on specific measures, involving several MAs and MSs in this debate at the same time. As mentioned above, the methodological complexity of these evaluations has to be overcome with improved planning of CIEs at national and international level.

Considering "soft outcomes"

In ESF interventions, CIEs generally measured effects with quantitative or binary (yes/no) variables, such as earnings or employment status. However, "soft outcomes", relating to the self-perception and capacity of the participants, are important measures of success in interventions. For instance, "employability", self-esteem, or acquired competences are pre-conditions to finding a job or to being active in the labour market. In many contexts, the analyses of soft outcomes would be even more useful than employment exit figures and would shine a light on the matching between interventions and individual needs.

CIEs can be applied to these soft outcomes on condition that the information is gathered before and after the intervention in both treatment and control groups in accordance with the theory of change of the intervention. These surveys must be planned early and involve all relevant actors, particularly the beneficiaries who are in direct and constant contact with the participants. For an example of a CIE assessing impacts on "soft outcomes" see box 21 below.

Box 21 An example of assessing effects on "soft outcomes" in Germany

Do Job creation schemes improve the social integration and well-being of the long-term unemployed? [105]

The study examines the effects of a job creation scheme targeted at a vulnerable group of people, long-term unemployed (LTU) who have been welfare claimants for at least four years and with health impairments or children, or both. The programme was operative from 2015 to 2018 and involved around 20,000 participants. The measure subsidised up to 36 months of regular work contracts for 30 working hours a week (mainly with public employers or charity organisations).

Since the subsidised employment measure, lasting up to 36 months, was explicitly aimed at promoting the social integration of the target groups, the authors tried to assess the impacts in these terms.

A rich dataset which integrated administrative data (the German integrated Employment Biographies) with a panel survey of programme participants and control individuals was used for the analysis. The panel survey enabled the collection of information on subjective measures of quality of life: life satisfaction, mental health, social belonging and social status. To collect information on these measures Likert scales were used, consistent with other surveys (such as the national survey PASS) in order to have comparable findings. The main steps of the survey were the following: treated and non-treated people were identified and matched in the administrative data; the resulting treatment and control groups were interviewed (three waves); the final dataset was built, cleaning the data and eliminating the cases with missing data. The

[105] See Ivanov B. et al., 2020.

sample was composed of 2,531 matched pairs in phase 1; 1,191 in phase 2, and 450 in phase 3 (out of about a total sample of 62,000 people, out of which 12,400 treated and 49,600 controls).

Relying on propensity score matching and measuring the effects at 7, 18 and 29 months after entry into the programme, the authors found that the interventions had positive effects on the measures of well-being, but to different degrees: for example, life satisfaction increased substantially while social status improved only moderately. However, the effects tend to decline over the course of the programmes and this is potentially explained by an increase in both the number of participants leaving the programme and control individuals finding a job. For the most vulnerable people positive effects were higher.

CIE integrated with TBE

As underlined at the beginning of this Guide, CIE is a powerful analytical instrument to assess “what” the ESF+ intervention has produced. To understand “how” or “why” the measured effects have been produced, it is necessary to use other instruments.

In some cases, where the interventions are well known and have already been evaluated in depth, a relatively limited analysis of the implementation and some interviews with the beneficiaries are sufficient to identify the main elements at the root of the measured effects. In more complex, less known or innovative interventions it is frequently necessary to integrate the CIE with other evaluations.

An implementation, or process, evaluation completed before the CIE may highlight how the intervention proceeded, the problems in execution and how the different actors collaborated to make it successful. The findings of this evaluation can feed the CIE, indicating which effects should be investigated, how the control group should be composed in relation to the participation characteristics, and the right timing for a CIE.

A Theory-Based Evaluation (TBE) may be carried out in parallel to the CIE. The two evaluations could mutually feed one another: the TBE provides explanations relating to the contextual, social and individual mechanisms that influenced effects and made them possible, while the CIE calculates net effects and clarifies to what extent the intervention produced significant outcomes. In an impact evaluation, the integration of CIE and TBE is recommended to provide a comprehensive overview of effects and their causes.

Glossaries

Acronyms

ALMP	Active labour market policy
CBA	Cost-benefit analysis
CIE	Counterfactual impact evaluation
CPR	Common Provision Regulation
DG EMPL	Directorate General for Employment, Social Affairs and Inclusion
DG REGIO	Directorate General for Regional Policy
DID	Difference-in-difference/s
EC	European Commission
EP	Evaluation Plan
ERDF	European Regional Development Fund
ESF	European Social Fund
EU	European Union
GDPR	General Data Protection Regulation (EU) 2016/679
IB	Intermediate body/ies
IEB	Integrated Employment Biographies
IV	Instrumental variable
JLD	Jobseeker longitudinal dataset
LFS	Labour Force Survey
MA	Managing Authority/ies
MS	Member State/s
NGOs	Non-governmental Organisations
OP	Operational programme/s
PES	Public Employment Service/s
PSM	Propensity score matching
RCT	Randomised controlled trial
RDD	Regression discontinuity design
SMEs	Small and medium sized enterprises
YEI	Youth Employment Initiative

Definitions

Term	Definition
Baseline data	Data on variables measured prior to a unit (individual or enterprise) being exposed to an intervention. In many cases pre-treatment measures of intervention results will be collected for both treatment and control groups.
Beneficiary	According to Council Regulation 1060/2021 (CPR), art.2, a beneficiary is a “a public or private body, an entity with or without legal personality, or a natural person, responsible for initiating or both initiating and implementing operations”. In the context of financial instruments (sometimes used for example for micro credits for self-employment in the ESF+ context), a beneficiary is the body that implements the holding fund or, where there is no holding fund structure, the body that implements the specific fund or, where the managing authority manages the financial instrument, the managing authority.
Control group	A group of persons, enterprises or other units, that is as similar as possible to the treatment group, but who remain untreated, and from which counterfactual estimates of results are obtained. Strictly speaking, the term "control group" refers to experimental settings (see "Randomisation" below), and the term "comparison group" refers to quasi-experimental settings, but in practice the two are used interchangeably.
Counterfactual analysis	A comparison between what actually happened and what would have happened in the absence of the intervention, in terms of the results. As the difference between actual and counterfactual results defines the causal effect of the intervention, the counterfactual analysis encompasses all approaches aiming to assess the proportion of observed change which can be attributed to the evaluated intervention.
Difference-in-differences (DID)	In its simplest form the difference in a result before and after treatment in a control group is subtracted from the same difference observed among a treated group in order to obtain an estimate of an intervention’s impact. Impacts calculated on the basis of difference-in-differences are usually derived within a regression framework.
Effectiveness	Refers to ‘achievement of objectives’ and is evaluated by comparing what has been obtained with what had been planned (or with a baseline situation) or by comparing what is observed after the action has taken place with what would have happened without the action (counterfactual situation).
Efficiency	Efficiency is defined as obtaining a given output at the minimum cost or, equivalently, as maximising output for a given level of resources. It can be established through cost-benefit or cost-effectiveness analysis.
Evaluation plan	According to Council Regulation 2021/1060 (CPR), art.44, the Member State or managing authority shall draw up an evaluation plan (EP) which may cover more than one programme and which should be submitted to the monitoring committee no later than one year after the decision approving the programme.
External evaluation	Evaluation conducted by an external and independent evaluator on the basis of a tendering procedure.
Impact	In the context of CIE, impacts refer to net effects, defined as the difference between average treatment and counterfactual results. For the purpose of this Guide, the term "impacts" is used interchangeably with "net effects".
Counterfactual impact evaluation	A type of impact evaluation that attempts to identify the causal effects of interventions through estimating average counterfactual results and subtracting these from average observed results among treated units. Estimates of counterfactual results are typically obtained from control groups carefully selected to be as similar as possible to the treated group.
Instrumental variable approach (IV)	In the application of this methods the selection into treatment should be at least partially determined by an exogenous factor (or instrument) which is unrelated to results other than through the treatment. Thus, the exogenous factor influences participation, but not directly the results.
Internal evaluation	Evaluation conducted internally, i.e., directly commissioned from an independent public institution or unit (from the MA or IB) without a tendering process or in the form of an extended monitoring and analysis process.

Term	Definition
Interventions	Refer generally to operations in ESF Operational Programmes or to projects co-financed by ESF.
Matching	It is a method in which intervention and control samples are matched to each other on the basis of their observed characteristics.
Non-randomised or quasi-experimental design	Approaches to counterfactual impact evaluation where control groups are constructed using methods other than randomisation.
Outcome	The likely or achieved short-term and medium-term effects of an intervention's outputs[106]. Outcome is similar to "result", but is used more often in impact evaluations.
Output	An output is considered everything that is directly produced/supplied through the implementation of an ESF operation, measured in physical or monetary units. Outputs are measured mainly in terms of number of supported people, supported entities, provided goods and services and implemented projects.
Participants	Participant is a natural person benefitting directly from an operation, without being responsible for initiating or for implementing such operation. In the ESF context it refers to people supported by ESF interventions.
Process evaluation	Process evaluation focuses on programme implementation, including, but not limited to, how services are delivered, differences between the intended population and the population served, access to the programme and management practices.
Propensity score matching (PSM)	PSM entails estimating a statistical model for the entire sample (treatment and potential controls) that yields an estimated propensity to participate for each individual or firm, regardless of whether they actually participated or not. Treated individuals or firms are then matched either to one untreated individual or firm, or to many untreated individuals or firms on the basis of the propensity score.
Randomisation	It is a method in which members of a target group are randomly assigned to a range of treatments or to control conditions. Randomisation ensures that groups are statistically equivalent in all aspects at the point they are randomised.
Regression discontinuity design (RDD)	This method may be undertaken when access to an intervention is determined by a cut-off point along a continuous rating, scale or measure. The approach makes use of the fact that those immediately around the cut-off point will be very similar to one another, but for the fact that those on one side of the cut-off point participate, whilst those on the other do not. Results for those above and below the cut-off can be compared to obtain an intervention's impact.
Relevance	Relevance refers to the appropriateness of the explicit objectives of an intervention with regard to the socio-economic problems the intervention is meant to solve[107].
Result	The effects on participants or entities brought about by an operation, for example in terms of their employment situation, earnings, scores in standardised education tests, profits, etc. The effects can be measured in the short term or in the longer term. In the ESF+ context, in measuring results indicators the short term is considered as immediately after participation (4 weeks), while the longer term is considered as six months after[108]. However, in the context of the impact analysis, longer term often means a longer period, 24 months or more after the interventions.
Treatment group	A group of persons, enterprises or other units, that benefit or are exposed to an intervention (this could be the offer of treatment or actual receipt thereof).

[106] See OECD, 2010.
[107] See European Commission, 2013.
[108] See European Commission, 2021a – and European Commission, 2018.

Bibliography

Ashenfelter, O. (1978). Estimating the effect of training programmes on earnings, Review of Economics and Statistics, 6, pages 47-57. https://www.jstor.org/stable/1924332?origin=crossref

Axdorph E, Egebark J., Lundström T., Özcan G. (2019) Effekter av förstärkta förmedlings-insatser för unga arbetssökande – Resultat från utvärderingen av Ung framtid. Arbetsförmedlingen

Baran J. et al. (2016). Badanie efektów wsparcia zrealizowanego na rzecz osób młodych w ramach Programu Operacyjnego Wiedza Edukacja Rozwój. I Raport Wskaźnikowy. http://files.evaluationhelpdesk.eu/Evaluations/PLE49.pdf

Baran J. et al. (2017). Badanie efektów wsparcia zrealizowanego na rzecz osób młodych w ramach Programu Operacyjnego Wiedza Edukacja Rozwój II Raport wskaźnikowy. http://files.evaluationhelpdesk.eu/Evaluations/PLE196.pdf

Baran J. et al. (2018). Badanie efektów wsparcia zrealizowanego na rzecz osób młodych w ramach Programu Operacyjnego Wiedza Edukacja Rozwój III Raport wskaźnikowy. http://files.evaluationhelpdesk.eu/Evaluations/PLE197.pdf

Baran J. et al. (2018a).Badanie efektów wsparcia zrealizowanego na rzecz osób młodych w ramach Programu Operacyjnego Wiedza Edukacja Rozwój. II Raport Tematyczny. http://files.evaluationhelpdesk.eu/Evaluations/PLE221.pdf

Bazzoli M., De Poli S., Rettore E., Schizzerotto A.(2018). Are Vocational Training Programmes Worth Their Cost? Evidence from a Cost-Benefit Analysis. In Politica Economica/Journal of Economic Policy Vol. XXXIV(3), pagg. 215-240. https://www.rivisteweb.it/doi/10.1429/92119

Bloom, H. S. (1995). Minimum detectable effects: A simple way to report the statistical power of experimental designs, Evaluation Review, 8(2), 225-246. https://cpb-us-e2.wpmucdn.com/sites.uci.edu/dist/1/1159/files/2021/03/Bloom-MDES-Eval-Rev-1995-Bloom.pdf

Bloom, H. S. (2009) Modern regression discontinuity analysis, MDRC Working Papers on Research Methodology, New York: MDRC. https://www.mdrc.org/publication/modern-regression-discontinuity-analysis

Boockmann B. et al. (2017). Evaluation des ESF-Bundesprogramms zur Eingliederung langzeitarbeitsloser Leistungsberechtigter nach dem SGB II auf dem allgemeinen Arbeitsmarkt. Im Auftrag des Bundesministeriums für Arbeit und Soziales. http://files.evaluationhelpdesk.eu/Evaluations/DEE189.pdf

Boockmann B. et al. (2018). Evaluation des ESF-Bundesprogramms zur Eingliederung langzeitarbeitsloser Leistungsberechtigter nach dem SGB II auf dem allgemeinen Arbeitsmarkt. Im Auftrag des Bundesministeriums für Arbeit und Soziales. http://files.evaluationhelpdesk.eu/Evaluations/DEE52.pdf

Boockmann B. et al. (2019). Evaluation des ESF-Bundesprogramms zur Eingliederung langzeitarbeitsloser Leistungsberechtigter nach dem SGB II auf dem allgemeinen Arbeitsmarkt. Im Auftrag des Bundesministeriums für Arbeit und Soziales. http://files.evaluationhelpdesk.eu/Evaluations/DEE94.pdf

Boockmann B. et al. (2021). Evaluation des ESF-Bundesprogramms zur Eingliederung langzeitarbeitsloser Leistungsberechtigter nach dem SGB II auf dem allgemeinen Arbeitsmarkt. Im Auftrag des Bundesministeriums für Arbeit und Soziales. http://files.evaluationhelpdesk.eu/Evaluations/DEE202.pdf

Borik V. et al. (2015). The net effects of graduate work experience and the promotion of self-employment. Technical Report. http://files.evaluationhelpdesk.eu/Evaluations/SKE8.pdf

Bratti M. et al. (2018). Vocational Training for Unemployed Youth in Latvia: Evidence from a Regression Discontinuity Design IZA DP No. 11870. http://ftp.iza.org/dp11870.pdf

Bratu C. et al. (2014). Knowledge gaps in evaluating labour market and social inclusion policies identified 39 ESF counterfactual impact evaluations in the 2007-2013 programming period. doi:10.2788/083390. https://op.europa.eu/en/publication-detail/-/publication/d0e73217-2a40-47af-a5c3-a700dcd47da3/language-en

Bredgaard, T. (2015). Evaluating what Works for whom in active Labour market policies. European Journal of Social Security, 17(4), 436-452. https://doi.org/10.1177/138826271501700403

Bryson, A., Dorsett, R. and Purdon, S. (2002) The use of propensity score matching in the evaluation of active labour market policies, Department for Work and Pensions, Working Paper Number 4. http://eprints.lse.ac.uk/4993/1/The_use_of_propensity_score_matching_in_the_evaluation_of_active_labour_market_policies.pdf

Caliendo M., Kopeinig S., (2008), Some practical Guidance for the implementation of propensity score matching. Journal of Economic Surveys; Volume 22, Issue 1 - Pages 31-72. https://doi.org/10.1111/j.1467-6419.2007.00527.x

Caliendo, M., Mahlstedt, R., & Mitnik, O. A. (2017). Unobservable, but unimportant? The relevance of usually unobserved variables for the evaluation of labour market policies. Labour Economics, 46, 14-25. https://doi.org/10.1016/j.labeco.2017.02.001

Card, D., Ibarraran, P. and Villa, J. M. (2011). Building in an evaluation component for active labour market programs: a practitioner's guide, Discussion Paper No. 6085, Bonn, Germany: IZA. http://ftp.iza.org/dp6085.pdf

Card D., Kluve J., Weber A. (2017). What works? A meta-analysis of recent active labor market program evaluations. Journal of the European Economic Association, Volume 16, Issue 3, June 2018, Pages 894–931, https://doi.org/10.1093/jeea/jvx028

Centre for Disease Control and Prevention CDC (2013). Good Evaluation Questions: Checklist to Help Focus Your Evaluation. Department of Health & Human Services - USA. http://www.cdc.gov/asthma/program_eval/AssessingEvaluationQuestionChecklist.pdf

Danmarks Statistik (2017). Effektmåling og monitorering 2016 - strukturfondsindsatsen i Syddanmark. http://files.evaluationhelpdesk.eu/Evaluations/DKE7.pdf

Danmarks Statistik (2018). Effektmåling af den Virksomhedsrettede Strukturfondsindsats 2007-2013. http://files.evaluationhelpdesk.eu/Evaluations/DKE10.pdf

Diaz J. M. et al. (2019). Evaluación del PO-FSE 2014/2020 del Principado de Asturias para el informe anual a presentar en 2019. http://files.evaluationhelpdesk.eu/Evaluations/ESE101.pdf

Ecorys, Ismeri Europa (2020) Study supporting the evaluation of ESF support to education and training (Thematic Objective 10) for the Commission, Directorate-General for Employment, Social Affairs and Inclusion. https://op.europa.eu/en/publication-detail/-/publication/d0c1a558-077d-11eb-a511-01aa75ed71a1/language-en/format-PDF/source-173162502

European Commission (2007): Indicative Guidelines on evaluation methods: evaluation during the programming period. Working paper no.5. DG Regional Policy. https://ec.europa.eu/regional_policy/sources/docoffic/2007/working/wd5_ongoing_en.pdf

European Commission (2013): EVALSED: The resource for the evaluation of Socio-Economic Development. Updated version. https://ec.europa.eu/regional_policy/sources/docgener/evaluation/guide/guide_evalsed.pdf

European Commission (2017) Better Regulation Toolbox, Tool #47 on *Evaluation criteria and questions* https://ec.europa.eu/info/sites/default/files/file_import/better-regulation-toolbox-

47_en_0.pdf

European Commission (2018) Monitoring and Evaluation of European Cohesion Policy European Social Fund. Guidance document. https://ec.europa.eu/sfc/en/system/files/ged/ESF%20monitoring%20and%20evaluation%20guidance.pdf

European Commission (2019) Advanced counterfactual evaluation methods. Guidance document. doi:10.2767/464242. https://ec.europa.eu/social/main.jsp?catId=738&langId=en&pubId=8254&furtherPubs=yes

European Commission (2020) Counterfactual impact evaluation of European Social Fund interventions in practice. Guidance document for Managing Authorities. doi:10.2767/55495. https://op.europa.eu/en/publication-detail/-/publication/f82c5fb8-bb40-11ea-811c-01aa75ed71a1/language-en

European Commission (2020) How to use administrative data for European Social Funds counterfactual impact evaluations. A step-by-step guide for managing authorities. https://op.europa.eu/en/publication-detail/-/publication/d96feed3-f30c-11ea-991b-01aa75ed71a1

European Commission (2021) Council Regulation n° 2021/1060 of the European Parliament and of the Council of 24 June 2021 laying down common provisions on the European Regional Development Fund, the European Social Fund Plus, the Cohesion Fund, the Just Transition Fund and the European Maritime, Fisheries and Aquaculture Fund and financial rules for those and for the Asylum, Migration and Integration Fund, the Internal Security Fund and the Instrument for Financial Support for Border Management and Visa Policy

European Commission (2021a): Common Indicators Toolbox Working document. June 2021. https://ec.europa.eu/sfc/en/system/files/2021/ged/toolbox_june_2021_final.pdf

Fondazione G. Brodolini, Metis GmbH, Applica, Ockham IPS (2020). Study for the Evaluation of ESF Support to Employment and Labour Mobility (Thematic Objective 8), for the European Commission, Directorate- General for Employment, Social Affairs and Inclusion. https://ec.europa.eu/social/BlobServlet?docId=22899&langId=en

Frolich, M (2004) Programme evaluation with multiple treatments, Journal of Economic Surveys, 18(2), pages 181-224. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.620.9209&rep=rep1&type=pdf

Guthmuller S., Meroni E.C., Evaluation of the higher education grant system for less privileged students in Portugal. JRC technical Report. Not yet published.

Hagglund, P. (2006). A description of three randomised experiments in Swedish labour market policy, Institute for Labour Market Policy Evaluation, Report 2006:4. https://www.ifau.se/globalassets/pdf/se/2006/r06-04.pdf

Heckman, J. J., Ichimura, H., Smith, J. and Todd, P. (1998) Characterizing selection bias using experimental data. In Econometrica, Vol. 66, No. 5. pp. 1017-1098. http://jenni.uchicago.edu/papers/Heckman_Ichimura_etal_1998_Econometrica_v66_n5_r.pdf

Hunger K. and Sattler K. (2017). Evaluationsbericht zum spezifischen Ziel A 1.1. im Rahmen der Evaluierung des Operationellen Programms des Europäischen Sozialfonds in Baden-Württemberg 2014–2020. http://files.evaluationhelpdesk.eu/Evaluations/DEE26.pdf

Holland, P. W. (1986). Statistics and Causal Inference, Journal of the American Statistical Association, 81 (396), 945-960 DOI: 10.1080/01621459.1986.10478354. http://people.umass.edu/~stanek/pdffiles/causal-holland.pdf

IAW Institut für Angewandte Wirtschaftsforschung - ISG Institut für Sozialforschung und Gesellschaftspolitik, GmbH (2015). Evaluation der Modellprojekte "Bürgerarbeit". Endbericht. http://files.evaluationhelpdesk.eu/Evaluations/DEE7.pdf

ICF, Cambridge Econometrics and Eurocentre (2020) Study supporting the 2020 evaluation of promoting social inclusion, combating poverty and any discrimination by the European Social Fund (Thematic Objective 9) for the European Commission, Directorate- General for Employment, Social Affairs and Inclusion. https://op.europa.eu/en/publication-detail/-/publication/8788ec85-2308-11eb-b57e-01aa75ed71a1

Indecon (2016). Indecon Evaluation of JobBridge Activation Programme. http://files.evaluationhelpdesk.eu/Evaluations/IEE2.pdf

Instytut Badań Strukturalnych, Imapp, IQS. (2015). Badanie efektów wsparcia zrealizowanego na rzecz osób młodych w ramach Programu Operacyjnego Wiedza Edukacja Rozwój. I Raport Tematyczny. http://files.evaluationhelpdesk.eu/Evaluations/PLE34.pdf

Ires Piemonte. (2019). Misure di sostegno all'imprenditorialitá. http://files.evaluationhelpdesk.eu/Evaluations/ITE61.pdf

Isfol (2016). Primo rapporto di valutazione del piano italiano Garanzia giovani. http://files.evaluationhelpdesk.eu/Evaluations/ITE13.pdf

Ismeri Europa (2018). I Rapporto tematico di valutazione – I risultati di PIPOL. http://files.evaluationhelpdesk.eu/Evaluations/ITE36.pdf

Ismeri Europa – Ecorys – Institute for Employment Studies. (2019). Pilot and feasibility study on the sustainability and effectiveness of results for European Social Fund participants using counterfactual impact evaluations. doi:10.2767/39339. https://op.europa.eu/it/publication-detail/-/publication/84cc9eb9-b33d-11e9-9d01-01aa75ed71a1

Ivanov B. et al., 2020, Do Job creation schemes improve the social integration and well-being of the long-term unemployed?, Labour Economics. https://doi.org/10.1016/j.labeco.2020.101836.

Kalinowski H. (2020). Badanie efektów wsparcia zrealizowanego na rzecz osób młodych w ramach Programu Operacyjnego Wiedza Edukacja Rozwój V Raport wskaźnikowy. http://files.evaluationhelpdesk.eu/Evaluations/PLE296.pdf

Kalinowski H. et al., 2020 Badanie efektów wsparcia zrealizowanego na rzecz osób młodych w ramach Programu Operacyjnego Wiedza Edukacja Rozwój Raport końcowy. http://files.evaluationhelpdesk.eu/Evaluations/PLE339.pdf

Kluve, J., Lehmann, H., & Schmidt, C. M. (2008). Disentangling treatment effects of Active Labor Market Policies: The role of labor force status sequences. Labour Economics, 15(6), 1270-1295. https://doi.org/10.1016/j.labeco.2007.12.002

Krug, G and Stephan, G. (2011) Is contracting-out intensified placement services more effective than in-house production? Evidence from a randomized field experiment, LASER Discussion Papers - Paper No. 5, http://doku.iab.de/externe/2011/k110912303.pdf

Lammers M. and Kok L. (2021). Are active labor market policies (cost-)effective in the long run? Evidence from the Netherlands. In Empirical Economics 60:1719–1746. https://doi.org/10.1007/s00181-019-01812-3

Lang, J. (2021) Employment effects of language training for unemployed immigrants. J Popul Econ (2021). https://doi.org/10.1007/s00148-021-00832-7

Małgorzata Z. et al. (2020). Ocena wpływu wsparcia RPOWP na popularyzację szkolnictwa zawodowego w województwie podlaskim. Raport końcowy. http://files.evaluationhelpdesk.eu/Evaluations/PLE359.pdf

Martini, A. (2009). Counterfactual impact evaluation: what it can (and cannot) do for cohesion policy, prepared for the 6th European Conference on Evaluation of Cohesion Policy, Warsaw, November 30th.

https://ec.europa.eu/regional_policy/archive/conferences/evaluation2009/abstracts/martini.doc

Morgan, S. L. and Winship, C. (2014) Counterfactual and causal inference: Methods and principles for social research. 2nd edition. Cambridge and New York: Cambridge University Press.

OECD (2010), Glossary of Key Terms in Evaluation and Results Based Management, Paris. https://www.oecd.org/dac/evaluation/2754804.pdf

Openfield (2019). Analiza skuteczności i efektywności dotacji na założenie działalności gospodarczej udzielonych wramach 8 osi priorytetowej RPO WM - komponent 3. http://files.evaluationhelpdesk.eu/Evaluations/PLE312.pdf

Palczyńska M. et al. (2019) Badanie efektów wsparcia zrealizowanego na rzecz osób młodych w ramach Programu Operacyjnego Wiedza Edukacja Rozwój IV Raport wskaźnikowy. http://files.evaluationhelpdesk.eu/Evaluations/PLE274.pdf

Pomatto G. (2017). L'attuazione del Buono per Servizi al Lavoro nella Regione Piemonte. Ires Piemonte. http://files.evaluationhelpdesk.eu/evaluations/ITE55.pdf

Pomatto G. (2019). Buoni per servizi al lavoro nella Regione Piemonte: qualità percepita dai destinatari e meccanismi dell'attuazione. Ires Piemonte. http://files.evaluationhelpdesk.eu/evaluations/ITE77.pdf

Pompili M., Giorgetti I. (2020). Rapporto di Placement. Servizio di attività di valutazione del POR FSE 2014/2020 – Regione Marche. http://files.evaluationhelpdesk.eu/Evaluations/ITE115.pdf

Pompili M., Giorgetti I. (2020a). Rapporto tematico "Disoccupazione di lunga durata". Servizio di attività di valutazione del POR FSE 2014/2020 – Regione Marche. http://files.evaluationhelpdesk.eu/Evaluations/ITE214.pdf

Poy S. (2019). Gli effetti occupazionali del buono per servizi al lavoro nella Regione Piemonte: prime evidenze. Misura per disoccupati da almeno 6 mesi, anno 2017. Ires Piemonte. http://files.evaluationhelpdesk.eu/evaluations/ITE123.pdf

Poy S (2020). Nuove evidenze sull'effetto occupazionale del buono per servizi al lavoro. Target persone disoccupate da almeno 6 mesi. Ires Piemonte. http://files.evaluationhelpdesk.eu/evaluations/ITE245.pdf

Re-source Pracownia Badań i Doradztwa (2015). Ocena działań na rzecz wydłużenia aktywności zawodowej osób starszych w kontekście niekorzystnej sytuacji demograficznej w województwie lubelskim. http://files.evaluationhelpdesk.eu/evaluations/PLE126.pdf

Riccio J., Friedlander, D., Freedman S. (1994). GAIN: Benefits, Costs, and Three-Year Impacts of a Welfare-to-Work Program, MDRC, NYC https://www.mdrc.org/publication/gain-benefits-costs-and-three-year-impacts-welfare-work-program

Scheller F. and Seidel K. (2020). Zweiter Evaluationsbericht zum spezifischen Ziel A 1.1: Teilnehmer*innenperspektive und Wirkungsanalyse. im Rahmen der Evaluierung des Operationellen Programms des Europäischen Sozialfonds in Baden-Württemberg 2014–2020. http://files.evaluationhelpdesk.eu/Evaluations/DEE171.pdf

Shadish, W. R., Cook, T. D. and Campbell, D. T. (2002). Experimental and quasi-experimental designs for generalised causal inference, Boston, US: Houghton Mifflin Company.

Stern E. et al. (2012). Broadening the range of designs and methods for impact evaluations, Department for International Development of the UK, Working Paper 38. https://assets.publishing.service.gov.uk/media/57a08a6740f0b6497400059e/DFIDWorkingPaper38.pdf

WK Kellog Foundation (2004). Logic Model Development Guide. https://www.wkkf.org/resource-directory/resources/2004/01/logic-model-development-guide

Annexes

Annex 1. Further reading

The following are suggested reading for managing authority personnel interested in more detail around issues touched upon in this Guide. The literature on evaluation is vast. This list is intended to point to reliable major discussions that provide immediately useful information for CIE planning. After each citation a short description of most sources is provided.

General Evaluation

- Rossi, Peter H., Mark W. Lipsey, and Gary T. Henry. (2018). Evaluation: A Systematic Approach. 8th edition. Thousand Oaks, CA: SAGE Publications.

  The classic textbook on evaluation practice and methods. Includes methods and examples.

- Gertler, Paul J.; Martinez, Sebastian; Premand, Patrick; Rawlings, Laura B.; Vermeersch, Christel M. J. (2016). Impact Evaluation in Practice, Second Edition. Washington, DC: Inter-American Development Bank and World Bank. © World Bank. https://openknowledge.worldbank.org/handle/10986/25030 (Available in English, Portuguese, and Spanish.)

  Like the present Guide, this handbook begins with classical (RCT) evaluation and then considers alternatives. While written for programme managers in lower-income countries, the discussion is relevant and readily applicable in EU Member State context.

- White, H., & Raitzer, D. A. (2017). Impact evaluation of development interventions: A practical guide. Asian Development Bank. https://www.adb.org/sites/default/files/publication/392376/impact-evaluation-development-interventions-guide.pdf

  Like Gertler et al. (2016) mentioned above, this guidebook – while addressing impact evaluation of interventions in low- and middle-income countries – presents a comprehensive discussion of general CIE methods. More technical than Gertler et al. (2016), but nonetheless with many practical insights that could be of interest for ESF programme managers.

- European Commission (2019) Advanced counterfactual evaluation methods. Guidance document. doi:10.2767/464242. https://op.europa.eu/en/publication-detail/-/publication/11968bbb-fac9-11e9-8c1f-01aa75ed71a1/language-en

  The document presents recent and more advanced counterfactual impact evaluation methods, such as sequence analysis, dynamic matching and synthetic controls.

- Csillag Marton, Kreko Judi and Scharle Agota. (2020). Counterfactual evaluation of youth employment policies. Prepared under the “Youth Employment PartnerSHIP” project. http://yepartnership.ibs.org.pl/content/uploads/2021/02/Methodological-guide.pdf (Available in English, Spanish, Hungarian (translation in progress), Italian and Polish)

  This is a “step-by-step” introduction to the counterfactual evaluation of labour market policies for youth, with a focus on the use of administrative data. Issues are presented on the basis of practical problems encountered in four CIEs of subsidies for hiring youth in Spain, Hungary, Italy, and Poland.

- HM Treasury (United Kingdom) (2020). The Magenta Book: Guidance for evaluation. London: The Agency. https://www.gov.uk/government/publications/the-magenta-book

  The “Magenta” book provides detail on evaluation methodology. These documents are interesting as examples of internal government evaluation perspective.

Randomised controlled trials

- White H. (2013) An introduction to the use of randomised control trials to evaluate development interventions, Journal of Development Effectiveness, 5:1, 30-49, DOI: 10.1080/19439342.2013.764652 https://www.tandfonline.com/doi/pdf/10.1080/19439342.2013.764652
- White, H., Sabarwal S. & T. de Hoop, (2014). Randomized Controlled Trials (RCTs), Methodological Briefs: Impact Evaluation 7, UNICEF Office of Research, Florence. https://www.unicef-irc.org/publications/pdf/brief_7_randomized_controlled_trials_eng.pdf

  Two non-technical introductions to the logic of RCT, with a discussion of several designs and criticism of RCT.
- Glennerster, R., & Takavarasha, K., (2013). Running Randomized Evaluations: A Practical Guide, Princeton University Press, Princeton, NJ.

  The book is a step-by-step guide on how to design and implement RCTs in the field of social programmes. It relies on the concrete RCTs carried out by Abdul Latif Jameel Poverty Action Lab.

Difference-in-differences

- Card, David, Pablo Ibarrarán, and Juan Miguel Villa. (2011). Building in an Evaluation Component for Active Labor Market Programs: A Practitioner's Guide. IZA Discussion Paper No. 6085. Bonn: Forschungsinstitut zur Zukunft der Arbeit. http://ftp.iza.org/dp6085.pdf

  Contrasts Diff-in-Diff with RCT.
- Lechner, M. (2011) The Estimation of Causal Effects by Difference-in-Difference Methods. Foundations and Trends in Econometrics. Vol. 4, No. 3 (2010) 165–224 DOI: 10.1561/0800000014. https://michael-lechner.eu/ml_pdf/journals/2011_Lechner_DiD_2011_ECO%200403%20Lechner_darf%20aufs%20Netz.pdf

  This paper discusses the DID approach in depth as well as some of the major issues in applying it. Extensions of DID, such as non-linear applications and propensity score matching with DID, are also presented.
- Fredriksson A. and de Oliveira G. M. (2019) Impact evaluation using Difference-in-Differences. RAUSP Management Journal Vol. 54 No. 4, pp. 519-532. DOI 10.1108/RAUSP-05-2019-0112 https://www.emerald.com/insight/content/doi/10.1108/RAUSP-05-2019-0112/full/pdf?title=impact-evaluation-using-difference-in-differences

  An overview of the DID methods, with practical recommendations
- Card, David and Alan B. Krueger. (1994). Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania. American Economic Review, 84 (4), 774–775. https://davidcard.berkeley.edu/papers/njmin-aer.pdf

  The classic example of application of a difference-in-difference technique.

Instrumental variables

- Morgan, Stephen L., and Christopher Winship. (2014). Counterfactuals and Causal Inference: Methods and Principles for Social Research. 2nd edition. Cambridge and New York: Cambridge University Press.

  This is a somewhat technical review of CIE methods using sociological terminology. Chapter 9, "Instrumental Variable Estimators of Causal Effects" (pp. 291-324) provides an overview of the logic of and procedures for IV estimation.
- Kuhn, Andreas, Jean-Philippe Wuellrich, and Josef Zweimüller. 2010. Fatal Attraction?

Access to Early Retirement and Mortality. IZA Discussion Paper No.5160. Bonn: Forschungsinstitut zur Zukunft der Arbeit. http://ftp.iza.org/dp5160.pdf

Uses regional variation in change in retirement age in Austria as instrumental variable in a study of the effect of early retirement on worker health.

- Galiani, S., Rossi, M. A., & Schargrodsky, E. (2011). Conscription and crime: Evidence from the Argentine draft lottery. *American Economic Journal: Applied Economics*. https://www.aeaweb.org/articles?id=10.1257/app.3.2.119

  Innovative and illustrative use of a randomised draft lottery (for conscription into military service) as an instrumental variable. Very readable study, prototypical for a line of research that uses lottery procedures as instruments.

Matching

- Heinrich, Carolyn, Alessandro Maffioli and Gonzalo Vázquez. 2010. A Primer for Applying Propensity Score Matching. Impact-Evaluation Guidelines Technical Notes No. IDB-TN-161. Washington: Inter-American Development Bank. https://publications.iadb.org/publications/english/document/A-Primer-for-Applying-Propensity-Score-Matching.pdf

  Like the regression discontinuity guide below, this is written to benefit knowledgeable evaluation managers.

- Caliendo M., Kopeinig S., (2008), Some practical Guidance for the implementation of propensity score matching. Journal of Economic Surveys; Volume 22, Issue 1 - Pages 31-72. https://doi.org/10.1111/j.1467-6419.2007.00527.x

  A classical text on the choices to be made in the implementation of PSM (in terms of estimation, matching algorithms, quality assessment of the matching, sensitivity of estimated treatment effects).

- Harris, H. and Horst, S. J. (2016) A Brief Guide to Decisions at Each Step of the Propensity Score Matching Process, Practical Assessment, Research, and Evaluation: Vol. 21, Article 4. DOI: https://doi.org/10.7275/yq7r-4820. Available at: https://scholarworks.umass.edu/pare/vol21/iss1/4.

  Similar to the previous article.

Regression discontinuity design

- Jacob, Robin, Pei Zhu, Marie-Andrée Somers, and Howard Bloom. (2012). A Practical Guide to Regression Discontinuity. New York: MDRC. https://www.mdrc.org/sites/default/files/RDD%20Guide_Full%20rev%202016_0.pdf

  Exceptionally accessible and thorough discussion of recession discontinuity methodology that includes a carefully selected bibliography

- Lee, D. S. and Lemieux T. (2009), Regression discontinuity designs in economics, NBER Working Paper No. 14723, National Bureau of Economic Research, Cambridge, MA, 2009. http://www.nber.org/papers/w14723.pdf?new_window=1

  The paper is a sort of "user guide" to Regression Discontinuity (RD) providing a description of the logic of the method and illustrating different ways of estimating RD designs.

Annex 2. Suggested CIE course outline

An introductory course in CIEs for MAs and officials might cover the following:

- Introduction to evaluation approaches: process evaluations (why they are important and how they relate to CIEs) and impact evaluations (counterfactual and theory-based approaches)
- What are CIEs? What evaluation questions can powerful CIEs answer? For which evaluation criteria do CIEs provide main evidence? Why are CIEs important?
- How do they work? (Treated and control groups, 'first and after' comparisons, inference principles)
- Overview of CIE methodologies (characteristics, strengths and weaknesses of each method)
    - Randomised control trial
    - Matching
    - Difference-in-difference(s)
    - Regression discontinuity design
    - Instrumental variable
    - Indication of other possible methods (e.g., techniques for spatial analyses)
- Data requirements
    - Requirements for treated and control groups
    - Possible sources (ESF monitoring, administrative datasets, surveys)
    - GDPR rules
- Overview of implementation steps:
    - Planning CIEs (evaluation plan, feasibility of a CIE)
    - Commissioning CIEs (terms of reference, selection criteria and methods)
    - Managing CIEs (verifying deliverables and interacting with the evaluator)
    - Dissemination of findings from CIEs (types of audience and communication channels).

A course structured as above should include practical examples to be developed in workshops and would last 2-3 days. One approach to delivering a course such as this would be to adopt a problem-based learning methodology and use a policy measure of the programme as the concrete case to develop.

Annex 3. Counterfactual Impact Evaluations – Examples mentioned in the Guide

Table 6 Characteristics of the CIEs mentioned in the Guide as examples

Country	Title - year	Policy evaluated and treatment period observed	Method	Treated and Control group – definition and size	Outcome variables	Period of observation	Data
Sweden	Effects of intensified support for young jobseekers - Results from the evaluation of Young Future (2019)	Intensified support of public employment service for jobseekers June 2017 - January 2018	Randomisation control trial (RCT)	TG: Young people 18-24 randomly assigned to receive the intensified support CG: Young people 18-24 randomly assigned to receive regular support TG: 2792 - CG: 1897	-proportion of unemployed -% of outflows to work -average number of days in unemployment	12 months after entering	-Public Employment registers for outcome variables -Survey of employment officers and intermediaries to measure the intensity of support provided
Italy (Marche)*	The Impacts of ESF interventions financed in 2014-2020 for long-term unemployed in the Marche region (2020)	Internships, work experiences, Job fellowships, training vouchers for unemployed (among which LTU) 2017 - July 2019	Propensity score matching (PSM)	TG: Long term unemployed treated CG: Long term unemployed in the period 2016-2018 registered at PESs TG: 526 for internships, 1058 for job fellowships, 236 for work experiences in the municipalities and 241 for training vouchers - CG: 77255	- probability of being employed at a certain time after the interventions - probability of being employed in an open-ended contract - number of days worked a period after the interventions	6,9,12,15 and 18 months after entering interventions	-ESF Monitoring data for treated group - Administrative data from PESs on people registered - Administrative data on labour contracts activated in the Region (Comunicazioni Obbligatorie - COB)
Poland (Lubelskie)*	Evaluation of measures for older workers in Lubelskie financed by the ESF in the 2007-2013 programming period (2015)	Several types of active labour market policies 2007-2013	PSM + Differences in Differences (DID)	TG: Sample of unemployed 50+ who received support CG: Unemployed 50+ who did not participate in any measure and who were registered in a Regional Employment Office at the time the treated group members took part in the evaluated programme TG: 145 - CG: 67102	-probability to exit from the condition of unemployed registered	6 12, 18 months after entering interventions	- Administrative data from PESs
Latvia	Vocational Training for Unemployed Youth in Latvia: Evidence from a Regression Discontinuity Design (2018)	Vocational training targeted at NEETs 15-29 2014-2015	Regression Discontinuity Design (RDD)	TG: Unemployed who participated in the Vocational training within one year from the registration date CG: Unemployed registered between June 2013 to December 2015 not participating in any measure TG: 898 -CG: 10717	-probability of being employed at a certain time after the interventions - income	12-36 months after enrolment	- Administrative data from the Latvian State Employment Agency (SEA) - information on both participants and non-participants registered as unemployed at specific dates - Administrative data from the State Revenue Service (SRS) - information on Employment condition at different dates and on the income of individuals
Germany	Employment effects of language training for unemployed immigrants (2021)	Training language courses (+work experiences for some of the treated) for migrant jobseekers	Instrumental Variables (IV)	TG: migrant participants who started language training for professional purposes in 2014 CG: random sample of non-participants with at least one period of	-probability of being employed (Regular employment, Regular full-time employment, Regular employment > 6 months, Employment with income above	24 months after entering interventions	- Integrated Employment Biographies (IEB) for the identification of both groups and for outcome variables - Data collected retrospectively

Country	Title - year	Policy evaluated and treatment period observed	Method	Treated and Control group – definition and size	Outcome variables	Period of observation	Data
		2014		non-German citizenship not participating in any active labour market policy measures or integration course s TG: 8968 - CG: 26463	risk of poverty threshold, Skilled employment) - daily and cumulated income from regular employment - cumulated days in regular employment		during meetings with jobseekers and caseworkers (Werdegangshistorie, WGH), for more information on episodes of self-employment or parental leave and on education and employment abroad
Italy (Trento)	Are Vocational Training Programmes Worth Their Cost? Evidence from a Cost-Benefit Analysis (2018)	Vocational training courses for unemployed people 2010-2011	PSM	TG: unemployed participants in vocational courses, financed by provincial resources and by ESF CG: people recorded as unemployed at the starting date of individual training courses in the registers of the local PES TG: 818 (province courses), 114 (ESF courses) - CG: 17236 (related to province courses), 1152 (related to ESF courses)	-probability of being employed -earnings	3,6, 12, 18, 24, 36 months after entering interventions	-ESF Monitoring data for treated group - Administrative data from PESs on people registered - Administrative data on labour contracts activated in the Region (Comunicazioni Obbligatorie - COB) - Administrative data on tax returns from INPS - Administrative data from Education Department of the Province
Germany	Do Job creation schemes improve the social integration and well-being of the long-term unemployed? (2020)	Job creation scheme (subsidized jobs for up 36 months) for vulnerable people (LTU) 2015 - June 2017	PSM	TG: sample of participants who answered to the entire questionnaire CG: non-participants registered with PESs and with the same eligibility criteria of control group, who answered to the entire questionnaire TG: 2531 in wave 1, 1191 in wave 2, 450 in wave 3 - CG: as above	- probability of being employed - subjective measures of life satisfaction, mental health, social belonging, social status	7, 18, 29 months from the entering into the programme	- Integrated Employment Biographies (IEB) for the identification of both groups and for outcome variables - Survey for subjective outcome variables (Soft outcomes)
Spain (Asturias)*	Evaluation of the Asturias ESF OP (2019)	Teaching methods for groups of students at risk of early school leaving (second and third year of secondary school) 2016-2017	PSM	TG: students participating to the interventions CG: students comparable in terms of eligibility criteria TG: 1053 - CG: 3852	- probability of passing to the next school year	12 months after entering	- Administrative data on education: SAUCE database (Ministry of Education)
Portugal	Evaluation of the higher education grant system for less privileged students in Portugal (2020) - to be published	Grants for students in higher education coming from low-income families 2021-2018	RDD	TG: students who apply for the first academic year of Master or Bachelor CG: students not eligible since their income is above the established threshold TG: 130602 - CG: 25400	- probability of being enrolled at the end of first year - number of credits obtained at the end of first year - probability of graduating - number of years to graduate - final mark at graduation	12-48 months from the grant	- Administrative data on students applying for the grant (DGES) - Administrative data with information on academic career and progression (DGEEC)

Country	Title - year	Policy evaluated and treatment period observed	Method	Treated and Control group – definition and size	Outcome variables	Period of observation	Data
Poland (Podlaskie)*	Assessment of the impact on the popularisation of vocational education of support from the Podlaskie OP in 2014-2020 (2020)	ESF project "Good Profession - Great Life", aimed to promote vocational education and training and targeted at lower secondary schools 2017-2019	PSM	TG: sample of students who participated CG: students having similar characteristics to those in the treated group, but not participating in the project TG: 200 - CG: 384	- attendance at a technical school at the time of the study - attendance at a 1st degree trade school at the time of the study - attendance at either a technical or 1st degree trade school at the time of the study - applying to a high school, a technical school, a 1st degree trade school, a school providing vocational training	12-36 months after interventions	- Monitoring data - National population registry (PESEL) for selecting the control group - Survey for measuring the outcome variables
Italy (Piemonte)*	Employment effects of the vouchers for employment services scheme (Buono servizi lavoro – BSL financed by the Piemonte ESF OP, 2014-2020) (2019)	Vouchers for employment services (orientation + training + stage) for vulnerable unemployed with an unemployment spell of at least 6 months 2018	PSM	TG: unemployed treated CG: unemployed registered at PESs with similar characteristics (30 years old and unemployed for 6 months) TG: 8125 - CG: 130000	- probability of being employed - probability of being employed with an open-ended contract	6, 12, 16 months from interventions	-ESF Monitoring data for treated group - Administrative data from PESs on people registered - Administrative data on labour contracts activated in the Region (Comunicazioni Obbligatorie - COB)
Germany*	Evaluation of the programme for the integration of long-term unemployed in Germany in 2014-2020 (2019)	Integration measures for LTU for 24 months or more (subsidized jobs) 2015-2017	DID (Intention to treatment was estimated)	TG: LTU potentially eligible in the period of the programme (from 2015) CG: people who meet the eligible criteria prior to the implementation of the programme (2010-2012) TG and CG: 134,515 people for 237,874 episodes of unemployment (35% for TG and 65% for CG)	- probability of being employed (employment with social contributions)	24 months from interventions	- Administrative monitoring data form Federal Office for Administration (ZUWES database) - Integrated Employment Biographies (IEB) for outcome variables
Germany (Baden-Württemberg)*	Second evaluation report of Specific Objective A1.1 under the Baden-Württemberg ESF OP, 2014-2020 (2020)	Measures to integrate unemployed, especially long-term unemployed and other vulnerable people March 2016-Decembre 2017	PSM	TG: participant identified in the IEB database CG: people who meet the eligible criteria in the same period TG: 1578 (out of total 1800) - CG: na	- probability of being employed (employment with social contributions) - probability of remaining on unemployment benefit	15 months from interventions	- Monitoring data - Integrated Employment Biographies (IEB) for outcome variables

Note: TG=Treated group, CG=control group
Note: * means evaluations identified in Helpdesk.

GETTING IN TOUCH WITH THE EU

In person

All over the European Union there are hundreds of Europe Direct information centres. You can find the address of the centre nearest you at: https://europa.eu/european-union/contact_en

On the phone or by email

Europe Direct is a service that answers your questions about the European Union. You can contact this service:

- by freephone: 00 800 6 7 8 9 10 11 (certain operators may charge for these calls),
- at the following standard number: +32 22999696 or
- by email via: https://europa.eu/european-union/contact_en

FINDING INFORMATION ABOUT THE EU

Online

Information about the European Union in all the official languages of the EU is available on the Europa website at: https://europa.eu/european-union/index_en

EU publications

You can download or order free and priced EU publications at: https://op.europa.eu/en/publications. Multiple copies of free publications may be obtained by contacting Europe Direct or your local information centre (see https://europa.eu/european-union/contact_en).

EU law and related documents

For access to legal information from the EU, including all EU law since 1952 in all the official language versions, go to EUR-Lex at: http://eur-lex.europa.eu

Open data from the EU

The EU Open Data Portal (http://data.europa.eu/euodp/en) provides access to datasets from the EU. Data can be downloaded and reused for free, for both commercial and non-commercial purposes.

Publications Office
of the European Union