Impact Evaluations Framework for the Spanish Ministry of Labour and Social Economy and Ministry of Inclusion, Social Security and Migrations

November 2020

Impact Evaluations Framework for the Spanish Ministry of Labour and Social Economy and Ministry of Inclusion, Social Security and Migrations

CONTRACT NO SRSS/S2019/036

This work was carried out with funding by the European Union via the Structural Reform Support Programme and in cooperation with the European Commission's Directorate-General for Structural Reform Support. This document has been shortened from the version presented to the Spanish authorities to present a general framework for impact evaluation that may be more broadly applicable to other countries.

Table of contents

Glossary

Term	Description
Activities	Actions taken or work performed within a programme to transform input into outputs.
Assumptions	Conditions to realise the Theory of Change that are or that are not within the control of the programme.
Average treatment effect	The effect of a programme across those that participated as well as did not participate in the programme.
Average treatment effect on the treated (ATT)	The effect of participating in a programme for those who participated in the programme.
Beneficiaries	Sub-group among the target group, who will (directly or indirectly) benefit from the outcomes of a programme. Usually considered the treatment group.
Comparison group	In a counterfactual impact evaluation, the group to which the effect on the treatment group is compared to evaluate the impact of the intervention.
Control group	Comparison group in experimental evaluation. Generally, the subgroup among the eligible population that has been randomly selected not to participate but to serve as a comparison for treatment group (Note that this is not necessarily equivalent to the whole remaining eligible group).
Counterfactual impact evaluation	A method to assess whether a policy produces the effects expected by the policymakers (i.e. the policy outcomes). The method involves comparing the expected outcomes for two groups i) those, who benefitted from a policy or programme (the "treatment group"), with ii) those, who did not benefit from the policy, but are otherwise similar to the treatment group (the "comparison/control group"). The comparison group provides information on "what would have happened to the participants in a policy in case they had not participated in it".
Counterfactual outcomes	The outcomes of individuals in the treatment group had they not received the offer to participate in the programme.
Effect	Change in the level of outcomes for each individual in the sample. Since the intervention can result in a change in the outcomes of non-beneficiaries (e.g. spill-overs) the effect can be defined for treatment and comparison groups.
Eligible group	Group of people eligible for the programme.
Evaluation	The systematic and objective assessment of an on-going or completed programme, its design, implementation and results. Quantitative evaluations analyse information (e.g. survey or monitoring data) to make judgements about benefits of programme or intervention.
External validity	An evaluation is externally valid if the evaluation sample accurately represents the population of eligible units. The results of the evaluation can then be generalized to the population of eligible units.
Impact (net effect)	Difference in the effect between treatment and comparison groups. Can be influenced by external factors and are typically achieved over a longer period of time.
Indicator	The performance standard to be reached to achieve an objective.
Inputs	Resources at the disposal of the programme, including staff and budget.
Intention-to-treat effect (ITT)	The ITT defines the impact of "offering" to participate in the programme (e.g. come to the initial information session). ITT analysis includes every subject who is randomised according to randomised treatment assignment. The ITT effect hence represents an estimate of the efficacy of the program, if it were rolled out to the entire eligible population as is (assuming that a similar share of eligible subjects may not comply as well).
Internal validity	An evaluation is internally valid if it provides an accurate estimate of the counterfactual through a valid comparison group.
Monitoring	Ongoing collection and analysis of data about the inputs, activities, outputs and outcomes of a programme or intervention.
Non-complier group (no-shows and drop-outs)	Typically, non-compliers includes all who are not adhering to their initial (randomised) treatment allocation. This includes individuals in the treatment group that do not participate (no-shows and drop-outs). It also includes individuals from the control group who participate in the programme (cross-overs).
Observables	Characteristics that are observable for the evaluator, i.e. for which the evaluator has data.
Outcome	Level of specific indicator for each individual in the full sample. Usually considered as the benefits of provided good or service to the beneficiaries of the intervention. Only achieved if the beneficiaries make use/profit from the outputs of a programme.
Output	Set of goods or services provided by an intervention. Part of the results of a programme.
Participant group	The subgroup of the treatment group that participates in the entire programme (e.g. excluding non-compliers and dropouts).

Result	Overall term for the outputs, outcomes and impacts of a programme that relate to the programme purpose.
Results Chain	Tool for analysing and presenting the most important elements of the Theory of Change of a programme and their interrelationships. Also called logical framework (log-frame), results chain, etc.
Selection bias	Occurs when the reasons for which an individual participates in an intervention are correlated with the (potential) outcomes this individual would observe under participation or non-participation. Ensuring that the estimated impact is free of selection bias is one of the major objectives and challenges for any impact evaluation
Theory of Change	Description of how an intervention is supposed to deliver the desired results. Hence the strategy for achieving the programme purpose, consisting of results, activities and means, and contributing to overall objectives. Also called programme theory or intervention logic.
Treatment	Term that refers to the policy, programme, intervention or event under study in an impact evaluation.
Treatment group	Subset of the eligible group who are selected for the "treatment" (e.g. by random selection from the eligible group). Often also called the beneficiary group.
Unobservables	Characteristics, which the evaluator cannot observe, i.e. characteristics for which the evaluator has no data.
Waitlisted group	Subset of the eligible group who are selected to receive later "treatment" (e.g. invitations) in case the initial treatment group does not reach the decided number of participants (due to non-compliance).

Abbreviations

ATE: Average treatment effect

ATT: Average treatment effect on the treated

CIE: Counterfactual impact evaluation

DID: Difference-in-differences design

ITT: Intention-to-treat effect

IV: Instrumental variable method

LATE: Local average treatment effect

MIS: Minimum Income Scheme

MISSM: Ministry of Inclusion, Social Security and Migration

MITRAMISS: Spanish Ministry of Labour, Migration and Social Security (in place until early 2020, before split to MITES and MISSM)

MITES: Ministry of Labour and Social Economy

RCT: Randomised controlled trial

RDD: Regression discontinuity design

SEPE: Public Employment Service

SGOPIP: General Secretariat for Inclusion and Social Welfare Objectives and Policies in MISSM

1. Introduction

The current Impact Evaluation Framework is the last output in the project that the OECD conducts together with the European Commission (DG REFORM), the Spanish Ministry of Labour and Social Economy (MITES) and the Spanish Ministry of Inclusion, Social Security and Migrations (MISSM). The objective of the project is to support the Spanish authorities in modernising their statistical and analytical system, in particular concerning data collection, analysis and communication.

During the first stages of the project, the MITES and MISSM were under one single ministry – the Ministry of Labour, Migration and Social Security (MITRAMISS). The first step of the project was to map the current situation regarding data management in MITRAMISS (so-called "as-is" analysis). The second step of the project proposed a new holistic approach for data management (so-called "to-be" analysis) for MITRAMISS.

In early 2020, MITRAMISS was divided to MITES and MISSM. For the last step of the project, both Ministries expressed interest to receive support on impact evaluation and building related analytical capacity. As such, this output closes the project by developing an Impact Evaluation Framework, tailored to the needs of both Ministries. The current Impact Evaluation Framework stands on the rich information received by the OECD for the mapping report (step 1 of the project) on topics such as the available data, infrastructure, analytical skills and practices. The framework is also in line with the proposed holistic approach for data management (step 2 of the project). More specifically, it stands on the following three building blocks:

- Promoting evidence-based policymaking
- State-of-the-art methods for conducting counterfactual impact evaluations
- Available (administrative) data for conducting counterfactual impact evaluations

In a follow-up project starting in Fall 2020, the OECD will continue to provide support on this topic by piloting the framework presented in this report. To facilitate the follow-up project, the current framework includes a roadmap for its implementation. As a result of these projects, the two ministries will have higher capacity, and in-depth knowledge to conduct impact evaluations in the future regularly by themselves and/or outsource these to third parties.

Content/structure of the report

Section 2 discusses the role of (counterfactual) impact evaluation[1] to inform evidence-based policymaking. Broadly, impact evaluations aim to assess changes in the well-being of individuals that can be attributed to a particular (public) intervention. The focus of impact evaluations is thus the *attribution* or *causal link* from an intervention to observed changes in outcomes. The key aim of counterfactual impact evaluations (CIEs) is to construct a comparison group that is able to credibly establish this causal link. This comparison group is used to provide an estimate of the *counterfactual*: "What outcomes would have

[1] The framework focusses on counterfactual impact evaluations throughout the chapters ("counterfactual impact evaluation" and "impact evaluation" are generally used interchangeably).

participants achieved had they not participated in the intervention under study?" CIEs thus address the "fundamental evaluation problem" that this situation is never directly observable.

Section 3 discusses key steps to ensure that impact evaluations are useful and used for evidence-based policy making. Already when designing the evaluation, evaluators should take steps to ensure that evaluation results will be picked up by policymakers and implementers. This involves setting evaluations in a broader framework in order to fulfil three objectives. First, ensuring that there is a demand for the evidence that is generated, i.e. that the questions of interest to policymakers are addressed. Second, ensuring that evidence is available at the right time and in the right format. Third, ensuring that the target audience (institutions and policymakers) are in the position as well as willing to adjust policy decision on the basis of this evidence.

Section 4 provides a general overview of the most common methods for CIEs. The methodological toolbox is introduced in the framework of the potential bias that can arise from (self-) selection of eligible individuals into the intervention under study. The chapter aims to provide a non-technical, short introduction to the most common methods and their underlying assumptions regarding selection into the intervention. To this end, the section clarifies first the key difference between experimental and observational studies – which is whether the evaluator has control over assignment to an intervention. These respective research designs are introduced in the two different sub-sections. First, the common experimental research designs (e.g. randomised assignment, randomised phase-in, and randomised encouragement) are presented. Second, the most common methods for observational studies are discussed. The latter sub-section differentiates between methods that assume selection is based entirely on characteristics observable to the evaluator, and methods that allow to account for selection on both observable and unobservable characteristics. The discussion on the econometric methods is accompanied by a discussion on which skills are needed to apply them and a suggestion on which technical tools (software) could be used.

Section 5 provides guidelines regarding data collection and linking to support CIEs. The wealth of administrative data collected by MISSM and MITES should be exploited to conduct evaluations, but this requires additional efforts to make the data available for research. In some instances, other sources of data may be needed to complement the administrative data, which requires to consider the data needs well in advance to address the current challenges of data sharing and linking. To structure the requirements for data to conduct a CIE, the ministries may want to consider the data needs to construct a treatment group, to obtain a comparison group, and to define the outcomes to evaluate. While constructing the treatment group essentially requires participation records, the data needs to obtain a comparison group depend on the CIE methodology. Experimental methods, as well as methods assuming selection on both observables and unobservables require much less data than methods assuming selection on observables only. At the same time, data availability also determines the choice of a method to be used. The outcomes to evaluate are determined by the objectives of the intervention, and more generally, by the mandate of MISSM and MITES. As such, MISSM may want to consider evaluating inclusiveness outcomes (labour market and social inclusion), wellbeing and poverty. MISSM may need to complement the administrative data collected with other sources of data to achieve these evaluation goals. MITES employment objective entails less ambitious data needs, as the information needed to define the labour market outcomes of interest is largely collected by the administrative records of MITES.

Section 6 presents a roadmap to apply the Impact Evaluation Framework, summarising the key activities presented in the previous sections of this document. The roadmap foresees four main phases to conduct a counterfactual impact evaluation: i) understanding the needs for counterfactual impact evaluations, ii) planning counterfactual impact evaluations, iii) implementing and managing impact evaluations, iv) disseminating results, ensuring policy uptake and managing knowledge. In a follow-up project starting in Fall 2020, the Impact Evaluation Framework will be piloted in both MITES and MISSM, following the roadmap presented in Section 6, its four main phases and the key activities in each of the phases.

2. What are counterfactual impact evaluations and why are they important?

Rigorous evaluation of public policies and programmes is a key step to inform policy making. Sound evidence on what works and for whom helps government to achieve strategic objectives and spending efficiency. Impact evaluations are one of the central parts of an evidence-based policy cycle. They serve as a foundation for greater accountability, innovation, and learning. Over the recent years, governments have started to acknowledge the importance of impact evaluations to inform policy. There are multiple drivers of this increasing demand, including budgetary commitments to increase cost-effectiveness; monitoring and evaluation requirements as part of specific funding arrangements (e.g., European Structural and Cohesion Funds); avoiding erosion of trust in public institutions; and new and complex policy objective and challenges.

2.1. Basic concepts: Policy intervention

This report defines a policy or a programme as an intervention. Broadly, (public) interventions are targeted to a specific population with the purpose of inducing a change in a defined state and/or behaviour (Loi and Rodrigues, 2012[1]).This definition highlights the four constitutive elements of a policy intervention:

1. **A target population**: a well-defined set of units (e.g. individuals, households, firms, geographic areas) upon which the intervention will operate at a particular time.[2]

   MITES: for example, jobseekers or jobseekers with specific obstacles for labour market inclusion, such as low digital skills.

   MISSM: for example, people under certain poverty threshold.

2. **An intervention**: an action, or a set of actions, whose effect on the outcome the evaluation wishes to assess relative to non-intervention.[3]

   MITES: for example, a digital tool supporting employment counsellors to counsel jobseekers or training on digital skills for jobseekers with low digital skills.

   MISSM: for example, Minimum Income Scheme accompanied by the dedicated policies, such as labour market inclusion itineraries.

3. **A set of participants and beneficiaries:** Units of the population that are exposed to the intervention are labelled as participants, while those who do not take part in the programme are labelled as non-participants. Units that are directly or indirectly affected by the intervention are labelled as beneficiaries.

[2] For simplicity of speech, the remainder of this document assumes the unit of observation is an individual.

[3] This report considers only interventions that consist of a single action rather than multiple alternative or subsequent actions (sub-interventions) and can be hence be represented by a simple participation vs. non-participation comparison.

MITES: for example, jobseekers who have been counselled by employment counsellors using a digital tool for counselling, or jobseekers with low digital skills who participated in a training on digital skills.

MISSM: for example, Minimum Income Scheme accompanied by the dedicated policies, such as labour market inclusion itineraries.

4. **An outcome variable**: an observable and measurable characteristic of the units of the population on which the intervention may have an effect.

   MITES: for example, labour market outcomes (employment, wage).

   MISSM: for example, labour market inclusion (employment, wage), social inclusion and wellbeing, poverty reduction.

2.2. Types of evaluation

Evaluations are a systematic and objective assessment of an on-going or completed project, programme or policy, its design, implementation and results (OECD, 2002[2]).[4] Evaluations assess the success of a programme or policy based on different *evaluation criteria*. The OECD evaluation criteria are commonly accepted as standard guidelines also beyond development assistance. They were updated in 2019 following a global consultation process. The revised guidelines describe six different criteria: Relevance, Coherence, Effectiveness, Efficiency, Impact, and Sustainability (OECD, 2020[3]). The evaluation criteria are typically the basis to define (more detailed) *evaluation questions* that should be answered by the evaluation.

Given the criteria that the evaluation seeks to address, the suitable *evaluation type* can be chosen. There are several types of evaluations and no standard classification exist. The most common evaluation types are:

- **Formative evaluations:** *Ex-ante* assessment whether a programme or intervention is feasible, appropriate, and acceptable before it is fully implemented. Mostly appropriate to assess the evaluation criteria "Relevance".

  Example for MITES: The evaluation could look at whether a digital tool for employment counsellors is likely to be needed by the counsellors and jobseekers, as well as understood, and accepted.

  Example for MISSM: An ex-ante evaluation whether the new Minimum Income Scheme (MIS) has the potential to reduce poverty, have no disincentive effects for becoming employed and how it fits in the framework of existing benefit schemes provided nationally and regionally

- **Process evaluation**: Determines whether programme activities have been implemented as intended. Conducted to assess the "Coherence" criteria.

  Example for MITES: The evaluation could look at how the digital tool for counsellors is used by the counsellors, whether the tool is used by the counsellors as intended by MITES.

  Example for MISSM: The evaluation could look at whether the MIS payments are made according to the guidelines as well as whether the labour market inclusion itineraries have been implemented according to the guidelines.

- **(Intermediate) outcome evaluation:** Measures intermediate programme effects in the target population by assessing the progress in the outcomes or outcome objectives that the programme is to achieve.

[4] This is a more focused, methodological definition of an impact evaluation, which is slightly different from a definition that focuses more the content of an impact evaluation, see for example Stern (2015[30]).

Example for MITES: The evaluation could look at whether the jobseekers who have been counselled using the digital tool change their behaviour, e.g. changes in job-search activity or enrolling to training programmes.

Example for MISSM: The evaluation could look at whether the MIS recipients who have followed the labour market inclusion itineraries have changed their behaviour regarding job-search activities.

- **Impact evaluation** (see Figure 1): Assesses programme effectiveness in achieving its ultimate goals.

  Example for MITES: The evaluation could look at whether the jobseekers who have been counselled using the digital tool become employed, earn higher wages, achieve sustainable employment, achieve better occupational match, etc. (after they have potentially changed their job search behaviour and/or participated in training programmes).

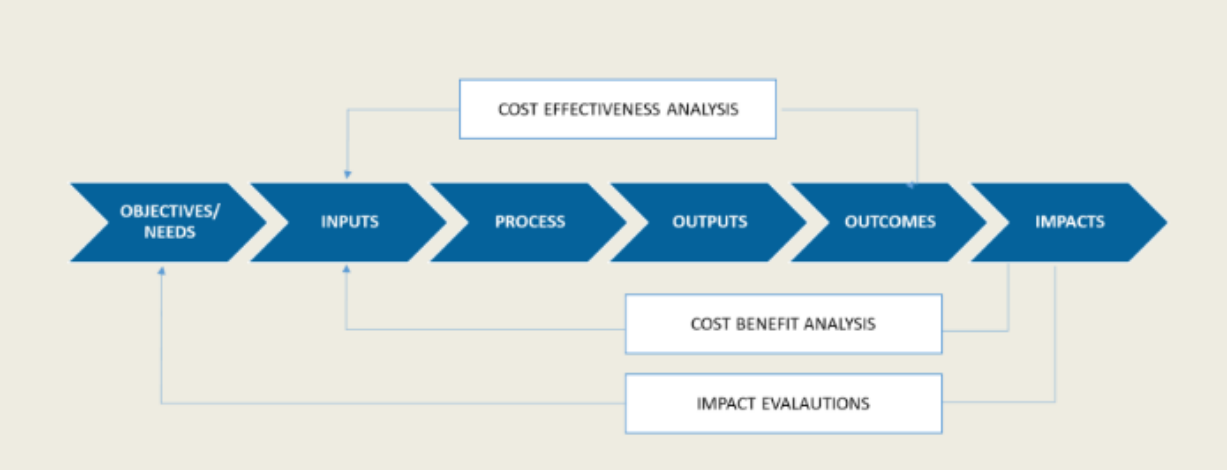
  Example for MISSM: The evaluation could look at whether the MIS recipients who have followed the labour market inclusion itineraries, have achieved a better labour market inclusion (earn higher wages, achieve sustainable employment, achieve better occupational match, etc.).

- **Cost-effectiveness and cost-benefit evaluation**: Examines the programme's outcomes (cost-effectiveness) or impacts (cost-benefit) in relation to the costs of implementing the programme and, if possible, the opportunity costs for beneficiaries (e.g. foregone earnings) as well as indirect costs on non-beneficiaries (e.g. negative externalities).

  Example for MITES: The evaluation could look at whether the benefits from higher employment rates and wages of jobseekers who were counselled using the digital tool are higher than the cost of developing and implementing the digital tool in employment offices.

  Example for MISSM: The evaluation could look at whether the benefits from higher labour market inclusion of MIS recipients are higher than the expenditures made on MIS.

Figure 1. Policy evaluation as part of the programme results chain



Source: (OECD, 2016[4])

Hence, Impact Evaluations are one specific type of evaluation. Impact evaluations assess whether the intervention being evaluated has effects on specific ultimate outcomes, irrespective of whether these effects are intended or unintended (see for example OECD (2006[5])). In many cases, an impact evaluation will be combined with other types of evaluations to address the full range of evaluation questions.

2.3. Why counterfactual impact evaluations?

One key issue for assessing the impact of an intervention is the so-called attribution problem. In most situations, observed outcomes depend on multiple factors and evolve over time independent of the intervention. Thus, changes in outcomes observed over the course of the intervention may have happened even in the absence of it. One example are coinciding (positive or negative) macroeconomic shocks that affect labour market outcomes of participants in a training programme. Attributing observed changes to the intervention requires isolating the particular contribution of the intervention and ensuring that causality runs from the intervention to the outcome (Leeuw and Vaessen, 2009[6]). Said differently, rather than assessing only which changes happened, impact evaluation is about attributing this change to the intervention.

Therefore, a valid assessment of the impact of an intervention requires establishing the "counterfactual": What would have been the level of outcomes in the absence of the intervention? Comparing the hypothetical outcomes that *would have occurred* without the intervention with *what has occurred* with the intervention provides an estimate of the impact. Defining a counterfactual is what differentiates counterfactual impact evaluations (CIEs) from purely monitoring levels or changes in outcomes before and after the intervention.[5] Only if a valid counterfactual can be established will the impact evaluation be able to credibly claim that the observed changes in outcomes can be attributed to the respective intervention.

Therefore, the core task in designing a CIE is to determine a method to estimate the counterfactual outcome. Of course, the counterfactual cannot be readily observed since one can only observe the world in *either* of the two scenarios: with or without the intervention. To overcome this problem, the typical approach is to determine a *comparison group* of individuals that is unaffected by the intervention, which therefore provides a credible basis to estimate the counterfactual. Section 3 gives an overview of some among the various methods that exist to identify such a comparison group.

Providing a valid and credible estimate of the impact should be the key building block for evidence-based policymaking. CIEs aim to provide evidence on "what works and what doesn't" (under what circumstances). Thus CIEs are key to deciding whether a policy or programme is successful or if it should be redesigned or discontinued altogether. If the impact of the programme is judged incorrectly, the resulting misleading policy advice may lead to the scaling up of inefficient or even harmful programmes. CIEs also provide the basis to conduct full economic evaluations either through cost-effectiveness or cost-benefit analyses. In this regard, CIEs also produce information that is relevant for accountability of government and other organizations. They produce knowledge about the benefits of a policy or programmes and what are the (financial) resources involved to reach these benefits.

[5] An assessment of whether the specific output or outcomes have been fulfilled is commonly referred to as "performance evaluation" (ILO, 2013[29]).

3. How to promote evidence-based policymaking

This section discusses key steps to ensure that impact evaluations are useful and used for evidence-based policy making. Ensuring that evaluation results will be picked up by policymakers and implementers starts already when designing an evaluation. This involves setting CIEs in a broader framework that aims to fulfil three objectives. First, ensuring that there is a demand for the evidence that is to be generated, i.e. that the questions of interest to policymakers are addressed. Second, ensuring that evidence is available at the right time and in the right format. Third, ensuring that the target audience (institutions and policymakers) are able and willing to adjust policy decision on the basis of this evidence.

The basic steps of conducting a CIE can be classified into four phases: i) understanding the needs for counterfactual impact evaluations, ii) planning counterfactual impact evaluations, iii) implementing and managing impact evaluations, iv) disseminating results, ensuring policy uptake and managing knowledge (the summarised key activities under each phase are presented also as a roadmap in Section 6).

3.1. What and when to evaluate? Understanding the needs for counterfactual impact evaluations

A first key step for deciding whether and what to evaluate is to assess demand for evidence by policymakers. In order to be useful for the audience, evaluations must answer relevant policy questions and bring actionable evidence to key stakeholders in a timely manner. During the very early stages of designing a CIE, all stakeholders of the evaluation (e.g. policymakers, programme managers …) should be engaged to identify the important policy questions. Ideally, a collaborative, iterative process with the evaluation team is ensued to judge the technical feasibility of answering those questions. Moreover, the dissemination of results should be agreed with stakeholders up front, clearly spelling out the evaluation objective, how to reach key audiences and a budget for conducting dissemination activities. This engagement process is a critical first step to influencing policy.

Two considerations are key to determine the feasibility of a CIE. First, evaluating an intervention should be technically feasible and provide the opportunity to deliver rigorous and valid results. Second, CIEs should only be conducted when they are likely to produce reliable and useful findings for policymaking. This can be determined as part of an evaluability assessment. The term “evaluability” describes “The extent to which an activity or project can be evaluated in a reliable and credible fashion” (OECD, 2010[7]). If there is not a clear theory of change of an intervention (a clear description of how an intervention is supposed to deliver the desired results), then findings from CIEs can be misleading.

Evaluation teams in the analysis units should systematically establish whether an intervention has a clear theory of change before its evaluation, and communicate it clearly to the stakeholders. A theory of change is the description of how an intervention is supposed to deliver the desired results, i.e. it is the strategy for achieving the programme purpose, consisting of results, activities and means, and contributing to overall objectives.

The theory of change can help specify the research questions, such as:

- Do we know how the intervention is expected to affect participants? And through which mechanisms?
- Are there clear outcomes to measure the effectiveness of the intervention? Are these reliable?
- Are there expected to be multiple interactions between different intervention components? How clearly are the expected interactions defined?

The role of analysis units and evaluation teams in the MITES and MISSM should be to promote evidence based policy, by explaining to policymakers the importance of evaluating these interventions with a clear theory of change, transmitting the challenges and limitations of evaluating certain interventions, and agreeing with them on the choice of interventions to evaluate and with which priority.

Evaluation teams in the analysis units play a crucial role in prioritising which interventions should be evaluated, given the limited resources to conduct CIEs. Some relevant questions to ask are:

- **Innovation**: is this evaluation testing a new and promising intervention?
- **Replicability**: can the intervention be scaled up or applied in a different setting? Can the conclusions from the CIE inform other interventions than the one evaluated?
- **Relevance**: does the intervention require substantial resources? Does it cover a large number of people?
- **Existing evidence**: is there no or little evidence about the effectiveness of the intervention from international evidence, or in a particular context?
- **Influence**: Is there strong political interest in the intervention? Will the results be used to inform key policy decisions?

As described in (Gertler et al., 2010[8]), the evaluation of interventions entailing a positive response to the previous questions should be prioritised.[6]

When prioritising the evaluation of certain policies, it is also essential to keep in mind the resources needed to conduct a CIE. Resource constraints may imply that not all policies can be evaluated, even if their evaluation would be valuable. Some key questions that need to be addressed are whether there are enough participants to be able to detect meaningful results (see Box 1), whether data on participants, a comparison group, and outcomes are being collected (see Section 5), and whether there are sufficient resources (time and financial) to conduct a CIE at a given moment.

[6] See also Kluve and Stöterau (2014[13]) for further guidance on a systematic selection of interventions to evaluate when assessing a portfolio of employment-related projects.

Box 1. Sample size, sample selection and statistical power

Sample selection

Sampling is the process of drawing units from a population of interest to estimate the characteristics of the entire population. In many cases, the target or eligible population may be too large to gather data on the entire population. This is especially the case for survey-based evaluations as resources for data collection usually increase with the sample size. To ensure that the sample is representative for the entire population, the process by which a sample is drawn is crucial. In practice, there are three main steps to draw a sample:

1. Clearly define the population of interest (i.e. the target group) and the unit of interest (individuals, firms…),
2. Identify a sampling frame: a comprehensive list of all individuals in the eligible population,
3. Draw as many units from the sampling frame as required to achieve the desired statistical power,
4. Choose a method to draw the sample (e.g. random sampling, stratified random sampling, cluster sampling).

Power calculations

Power calculations are a key step to determine *ex-ante* whether the foreseen impact evaluation design will be able to provide valid results.[7] Depending on the setup, this question can be addressed in two ways: (1) How large does the evaluation sample need to be in order to provide reliable estimates of a given expected programme impact? (2) How large does the programme impact has to be in order to be detectable with a given sample size (called the "minimum detectable effect size"). The first question is usually of relevance when the evaluation collects survey data and thus larger samples lead to higher evaluation costs. For evaluations based on administrative data, the second question is more relevant: The key rationale is to determine whether the eligible group is sufficiently large to detect the expected impacts. In case the intervention is assumed to deliver comparatively small (true) impacts, a larger sample size is required to ensure that the evaluation can reliably detect even potentially small impacts.

Power calculations should be conducted as part of the evaluation assessment phase. They are a key aspect to determine the programme can be conducted given the specific programme design and resources for evaluation. In many cases, such a simple power calculation will already provide a reliable rough estimate of the required sample size. Several software tools exist to perform power calculations and most data analysis programmes include prebuild packages (e.g. Stata's *sampsi* command). One common freely available software is Optimal Design. For basic (non-clustered) power calculations, an easily accessible online tool is called SWOG.

The cost of rigorous CIEs can vary significantly. CIEs' costs can increase when data need to be collected additionally (instead of relying on existing (administrative) data), for multiyear evaluations that require frequent follow-up, when the design of the CIE is complex or tests multiple hypothesis or when the CIE requires the involvement of several evaluators. Analysis units in the ministries can adjust these parameters to make more room in the budget to conduct CIEs.

[7] In more technical terms, the (statistical) power of an impact evaluation is the probability that it will detect a difference in the outcomes between the treatment and comparison groups, when in fact one exists (i.e. there is a low risk of *not* detecting the true *existing* programme impact).

Getting the timing right when conducting a CIE is key. When conducted too late, findings from the evaluation come too late to inform policy decisions. When conducted too early, the results may provide an inaccurate picture of the policy impact. Some questions to be asked are thus, is there an opportunity for the findings of the impact evaluation to have an influence? Has the policy been around for a sufficient time (and have sufficient take-up) to enable useful lessons to be extracted? If the evaluation was planned in advance, is the evaluation still relevant?

3.2. Planning counterfactual impact evaluations

Detailed planning of a CIE by the analysis units can ensure its success, lower its costs and minimize institutional disruptions. Similarly to public programmes, impact evaluations are often implemented with a given budget and their benefits should be judged in relation to the costs. In addition, impact evaluations often demand additional work effort by policy managers and staff, and may even require changes to the programme implementation. These can be reduced through a clear understanding of the programme operations. There are several elements of a CIE to scrutinize in the planning phase – including their data collection, CIE design and actors involved.

Good data are one of the main costs of impact evaluations, but these can be lowered with proper planning. Identifying requisite data sources early enough to have time to be granted access, or ensuring that there is an appropriate recordkeeping that can be integrated into the delivery of interventions, can facilitate CIEs at a lower cost than when relying on survey data. Moreover, data collection can often be integrated into existing procedures of the intervention. For example, additional data can be collected and inserted in the administrative database if these are deemed to be useful for a CIE in the future. Furthermore, it is important to estimate the needs for necessary sample size prior to conducting a CIE, see Box 1).

Whether the evaluation was planned before the programme started or during or after the programme began determines the range of potential methods. Planning CIEs ahead and integrating them in the programme design phase has many advantages. CIEs that are planned in advance offer more options for evaluation methods than when programme or policy has been rolled out. Oftentimes, experimental approaches are (more) difficult to implement once the programme has started (i.e. when there are not subsequent programme cohorts). In addition, planning ahead allows to collect baseline data in order to establish and test whether the evaluation design generated appropriate comparison groups.

Taking a forward-looking approach to policy evaluation may allow the design of the evaluations to be simplified. For instance, the implementation of policies can be modified in subtle ways that can simplify the design of the CIEs. For example, if the MITES or the MISSIM were to roll out the interventions to be evaluated geographically, non-treated regions or municipalities could act as a comparison group to treated ones. This would allow for a much simpler design of the CIE. Generally, if the intervention is implemented such that the requirements of CIEs are met, the need for a complex design and thus very strong analytical skills and extensive data can be lowered.

The planning phase should assess who is most qualified to design and implement the evaluation. The key question is whether to conduct the CIE internally, entirely contracted to external parties or a mix of both. Conducting CIEs internally within the respective public body (“in-house”) usually entails that only civil servants are involved in the evaluation. This has the advantage of limiting the need to develop contracts with external parties, legal agreements to access data, and can result in lower coordination costs by offering greater flexibility to mobilise researchers. Even if CIEs are conducted internally, however, the ministries may want to outsource some of their CIEs to maintain the independence of the policy evaluators, or to establish external quality insurance bodies.

Commissioning CIEs to an external evaluation team may be advantageous when the internal analytical skills are limited. The ministries may opt for having a smaller analysis unit in-house, charged with commissioning CIEs to external researchers (consultancy firms and other private companies, think tanks, universities etc.).[8] This can take the form of *ad hoc* collaborations with researchers or consultants, through one-off bids or a cohesive collaboration with the research community. In such cases, academic researchers can leverage knowledge on state of the art econometric techniques to public sector employees' expertise on institutional characteristics and administrative data (OECD, 2020[9]). Such arrangements do not need to be static, and can evolve towards more involvement of civil servants as they gain more experience in conducting CIEs. Note that contracting out CIEs requires still a certain level of analytical skills to be able to draft procurement documents and select the research proposals that would deliver evaluations of high quality (see Section 5.8). The potential evaluation/analysis teams in MISSM and MITES should get support from the departments dedicated to procurements when contracting out CIEs. The analysis teams should draft the objective and technical requirements for the CIE (what, when and how should be evaluated), but the general procurement process should be conducted by the experts in public procurement in Spain.

Going a step further, the ministries could facilitate access to administrative data, thus allowing researchers to engage in activities that can overlap with ministries' evaluation needs. This can represent a win-win situation, since the empirical research activity on MISSM and MITES policies can grow, thus building the evidence base for policymaking for "free" (OECD, 2020[9]). This requires a strong collaboration between the ministries and the research community, to ensure that the focus is aligned with the needs for evidence. While the practice of sharing data could be improved to some degree by the activities of the statistics/analysis units themselves, ideally the supporting IT infrastructure should be further developed as well. This would require a cooperation with the IT units and data protection delegates, as well as additional investments.

3.3. Managing an impact evaluation

Even well-designed and well-planned evaluations are not easy to implement. The implementation of the evaluation needs to be closely monitored by the analysis units to identify potential issues as early as possible. This is in particular the case for experimental study designs, since they are typically based on a direct connection to intervention design and may even entail changes from the regular procedure of the intervention. For example, the assignment process has to be integrated into the intervention design, processes have to be defined, and implementation guidelines have to be written, adaptions to the IT infrastructure might be necessary. Observational studies often involve less of an active management component, since they tend to be conducted after the intervention has been implemented. Thus, data analysis takes up a larger role than implementation of the design. However, other parts of managing an impact evaluation may still be necessary.

Conducting CIEs therefore requires constant management efforts from the analysis units, ideally following a Plan-Do-Check-Act (PDCA) management cycle. This implies that after designing the impact evaluation strategy as described in sections 3.1 and 3.2 (Plan), the strategy and more detailed plans have to be also implemented (Do), the implementation progress has to be constantly monitored (Check) and based on the monitoring results, actions have to be taken (Act). Such a management cycle would allow for reflecting on the process of the CIE afterwards and drawing lessons from this for future impact evaluations.

[8] Commissioning is optimal in some cases, for instance when data collection is necessary, since the ministries have limited capacity to conduct surveys in-house.

Implementation often involves a collaborative effort from policymakers, evaluators, programme managers and staff. Involvement of key stakeholders by the analysis units, close communication and collaboration is therefore a key aspect to ensure a successful outcome. In case experimental evaluations are conducted, managers and staff implementing the intervention will have to be particularly well informed about the goal, rationale and process of running an experimental CIE. This is particularly important in cases where staff will have to explain (or even justify) the procedure to potential participants. Most important is often to convey the usefulness of running a CIE to the programme implementers for their own work. Ideally, programme managers and staff should have a sense of ownership for evaluation and its success (e.g. in case of MITES and potentially evaluating the impact of a digital tool for counsellors, the employment counsellors in the employment offices should understand why it is important to evaluate the impact of this tool).

Planning and managing CIEs calls for setting up dedicated teams in MISSM and MITES assigned with managing impact evaluations, or some staff in each analysis unit tasked with these duties. The teams should be dedicated to day-to-day managerial tasks, commissioning external evaluations or coordinating in-house analytical activities, advising researchers and reacting to any issues that come up, including concerning data linking and sharing. The dedicated CIE teams should also take care of disseminating the results. One option for the ministries, particularly when most CIEs are contracted-out, is that the teams or staff in charge of analytical tasks would be also charged with the managerial tasks of CIEs. As more CIEs are conducted in-house, the ministries may want to identify dedicated staff to manage CIEs. These staff need to have analytical skills and a good knowledge of the data, although to a lesser extent than staff conducting the analytical part of CIEs.

3.4. Disseminating evaluation results to staff delivering the intervention and policymakers

Impact evaluations can only provide a benefit for policymaking if the results are disseminated to the intended audience. Generally, monitoring and evaluation of public policies can follow two different purposes (OECD, 2020[3]). First, **accountability** (transparency, communication) – i.e. to inform the stakeholders or general public whether the evaluated intervention achieved the desired results. Second, **learning** (understanding) – i.e. to inform staff delivering the intervention and policymakers how the design and delivery of interventions may be improved. The goal is to guide policy decisions so that scarce resources yield the highest social returns possible. While those policy decisions will be influenced by a range of factors, from the political economy to ideological positions, evaluations can support policymaking by providing a solid evidence base.

Full transparency throughout the entire evaluation cycle strengthens the trustworthiness of the evaluation and its potential for influencing policy. Impact evaluations tend to produce large volumes of information (e.g. decisions on the evaluation design, descriptive statistics and data sets, statistical code). While not all information is relevant to the stakeholders, it is crucial to document and publish all information throughout the evaluation cycle since, ultimately, the credibility of the evaluation results may hinge on the technical details. Ideally, the documentation should allow that the evaluation is entirely reproducible from the provided information. One way to ensure access to all material is to set up specialized websites or evaluation repositories that collect information on impact evaluations in a standardized fashion.[9]

[9] For example, a ready-made platform to share data, materials or code is provided by the Center for Open Science (https://www.cos.io/). An existing example is Harvard's Data-verse repository that contains data and statistical code underlying published research results from randomised experiments in the social sciences: https://dataverse.harvard.edu/dataverse/DFEEP

Findings from evaluations must be disseminated in a form that decision makers can easily access and use. While completeness and transparency are critical, most information users will not need to or want to delve into the details. It will be up to the evaluation team to distil a manageable set of key messages summarizing the most policy-relevant results and recommendations, and to communicate these messages consistently across audiences. Moreover, the communication should be tailored to the target audience(s). Each audience group (policymakers, staff delivering the programme, but also civil society and the media and programme participants – see the next subsection) will have different interests and background knowledge to understand the evaluation results.[10] The sequencing of dissemination activities may also be critical. For example, prior to public dissemination, an internal round of presentations and consultations should be conducted with programme staff and managers to avoid that premature results hurt a programme's reputation.

Ideally, the evaluation results should feed into action plans to change approaches, re-designed policies and guidelines for policy implementers. The task of the evaluation teams in MISSM and MITES would be to communicate clearly the results of the CIEs and make recommendations for change based on these results. While re-designing policies and up-dating guidelines for policy implementers would be the tasks for the units in charge of policy design in the ministries, the evaluation teams could be ideally involved in these processes, supporting the policymakers in translating the evaluation results to policy design.

3.5. Disseminating evaluation results to wider public and managing knowledge

After initial dissemination to key stakeholders, evaluation results should generally be made available to the civil society. Informing the public of the results of an evaluation through the media can play a key role in achieving accountability for public spending, building public support for the evaluation recommendations, and sustaining effective policies**.** This is particularly true of new and innovative policies where the outcome was initially uncertain or the subject of controversy in the policy debate. In addition to the broader public, evaluators may consider informing programme or evaluation participants specifically since they often spent considerable time providing information (e.g. answering surveys or participating in qualitative interviews).

The design and implementation of a communication strategy that ensures policy uptake can be highly complex and time-consuming task. Proper planning and budgeting of dissemination activities, so that the results of the evaluation reach their intended audiences quickly and effectively, is key to maximize the policy impact. Since communication is oftentimes not the key expertise of evaluators, the evaluation team may be supported by (internal or external) communication experts. Hence, the analysis units (or potential evaluation teams) in MISSM and MITES should be above all in terms of what to disseminate. However, the communication experts in the ministries should advise the analysis units in terms of how to disseminate – which channels to use, whom to target, how to phrase and communicate the key messages.

Effectively managing the knowledge generated from systematic CIEs is a key step to influence policy in the longer term. A single impact evaluation can only provide partial information, set in a specific context and often highly dependent on the specific implementation details of the policy. Ideally, multiple impact evaluations should be available to inform policy making on a broader question. As MITES and MISSM engage in conducting more CIEs of their policies, they have to ensure that the growing body of evidence is easy to locate. Evaluation results should be integrated into larger (public) repositories, such as these managed by the Danish Agency for Labour Market and Recruitment (https://www.jobeffekter.dk/en/), the US Department of Labor's Chief Evaluation Office (https://clear.dol.gov/), among many others (see the European Commission Database of national practices

[10] See Chapter 14 in Gertler et al. (2016[11]) how to tailor a communication strategy for different audience groups.

on European employment policies and measures for more examples). These repositories should consist of an easy-to-use and searchable gateway. Each evaluation could be accompanied by a short summary of the intervention and target population, methodology used, main results, and a transparent evaluation of the quality and trustfulness of the results as in the Danish repository. Information on the evaluation team should be available to ensure that users can contact them. CIEs of interventions, even if not conducted in Spain, could also be made available and summarised in the repository, thus allowing to synthesise the different existing evidence, as well as easily highlighting knowledge gaps and priority evaluation areas.

4. Micro-econometric methods to evaluate policy impact

Broadly, a counterfactual impact evaluation assesses the changes in the well-being of individuals that can be attributed to a particular (public) intervention. The focus of CIEs is thus the attribution or causal link from the intervention to observed changes in outcomes. To credibly establish this causal link, CIEs are typically based on quantitative data and micro-econometric statistical methods (i.e. analysing data on the individual or firm-level rather than macroeconomic data). A large methodological toolbox to conduct impact evaluations has been established, which can be broadly classified into methods for experimental (randomised) studies and observational studies. This section discusses the core concept underlying the most widely used methods. On this basis, the sections aim to provide an intuition of the key assumptions and approaches of the different approaches within the methodological toolbox.

4.1. The fundamental evaluation problem: the need to answer the questions of "what would have happened to the participants in a policy in case they had not participated in it"

The core goal of impact evaluations is to identify the causal effect of a programme or policy. The key evaluation question is thus: "What is the impact of the intervention on the outcome of interest?" This impact (causal net effect) is defined as the difference between the outcome of individuals affected by the intervention (the actual, observed situation) and the outcome that these same individuals would have experienced had the intervention not been implemented (the counterfactual). The fundamental evaluation problem is that – by definition – one cannot observe simultaneously the same individual in both scenarios. Either of the scenarios remains a hypothetical state for which the potential outcomes have to be simulated or estimated from other data available. The most common approach, discussed in this section, is to estimate the counterfactual from the observed outcomes of a comparison group.

A naïve attempt to measure an intervention's impact can result in misleading estimates. Two naïve approaches often come to mind: i) comparing the level of outcomes of the beneficiaries before and after the intervention (known as before-after, pre-post or reflexive comparisons); and ii) comparing outcomes of beneficiaries with outcomes of a (random sample) of individuals that are believed to be unaffected by the intervention. However, neither approach will credibly yield the true causal effect. In the first case, outcomes may be affected by other factor which changed simultaneously over time (thus one observes a "spurious correlation"). In the latter case, outcomes of beneficiaries and the (naïve) comparison group may have been different even in the absence of the intervention. This is particularly the case when individuals can choose to participate in the intervention (e.g. the incentives and motivations can differ considerably). This (self-) selection leads to the so-called selection bias that is commonly considered the core issue to address by a CIE. Defining a valid comparison group is therefore the key challenge of designing an impact evaluation.

4.2. The methodological toolbox for counterfactual impact evaluations

A large toolbox of methods exist to identify a valid comparison group for conducting a CIE. To organize this toolbox, a variety of different terms and classifications have been put forward in the literature over the recent years. This following classification is based on commonly used terminology while also providing the conceptual clarity to present the differences in the assumptions of each method.[11] As described in detail below, this classification is based on (i) whether the evaluator can control assignment to the intervention (experimental vs. observational studies) and (ii) within each of these two categories, the specific research design (or "methods"), see Table 1. It should be noted, however, that a variety of different classification systems of methods are encountered in the evaluation literature. For example, in some guidebooks, all research designs for observational studies are grouped together and referred to as either "Non-randomised studies", "Non-experimental methods" (Ruiz and Love, 2012[10]) or "Methods using non-experimental data". Impact evaluation methods that use a counterfactual, but are not based on randomised assignment of the intervention, are often called also "Quasi-experimental methods" (Gertler et al., 2016[11]).

The key aspect for classifying CIE methods concerns whether assignment to the intervention is under the control of the evaluator or not. If the evaluator has determined assignment, this is typically referred to as experimental studies (randomised controlled trial, RCT). In this case, the evaluator has the opportunity to randomly assign individuals from the target population (or a random subset of them) to treatment group and the comparison group prior the start of the intervention. This is often called the "gold standard" research design for policy evaluation. Since assignment is under control of the evaluator, in experimental studies the comparison group is often called the control group.[12] The key differences among research designs used in experimental studies regard the question how the (randomised) assignment is administered in practice (see Section 4.4).

Oftentimes, the evaluator cannot influence assignment to the intervention, e.g. because of ethical concerns, logistical constraints or timing of the evaluation. Rather, the evaluator simply observes the (externally allocated) participation in an intervention. In this case, the evaluator will have to resort to an observational study. The available methods for observational studies can be classified into two types of research designs – with the key difference between them being whether or not the respective method assumes selection only on observable characteristics or also on unobservable characteristics (see Table 1). In short, one can consider that these research designs differ in their approach how they construct a credible comparison group (c.f. Wooldridge (2009[12]), Kluve and Stöterau (2014[13])). The methods assuming selection on observables and unobservables construct a comparison group by exploiting events that exogenously change the probability to participate in the programme for a quasi-random subset of the target population. If such an external factor or event cannot be found, the evaluator has to assume that the available data covers all characteristics that drive selection into the treatment group. Methods that rely on this assumptions use statistical techniques based on data collected about observable characteristics to re-construct the counterfactual.

It is important to understand that the research design is distinct from the empirical model used to estimate the treatment effect. For example, a linear regression is an empirical model for covariate adjustment, whereas a "matching" research design can involve estimating a linear regression. This often leads to confusion about the terminology and concepts. Irrespective of the research design, the evaluator can choose different empirical models to analyse the data. Most commonly, researchers use linear regressions (e.g. ordinary least squares) or non-linear regressions (e.g. probit or logit model). The choice

[11] The classification is similar to that proposed in (e.g. Kleinbaum, Kupper and Morgenstern (1982[28]), (Costantini and Higginson (2007[27])

[12] This distinction not always (but increasingly) made in the literature. This note follows this distinction for conceptual clarity, but both terms are still encountered interchangeably in the literature.

of methods depends on the underlying outcome variable. An RCT can assess impacts on wage, in which case one could estimate an ordinary least squares regression or simply compare averages. An RCT can also assess impacts on the duration in unemployment – in which case one would use non-linear models (e.g. a duration model). Evaluation approaches are often confounded with empirical methods. But clarifying the difference is also important in practice. For example, matching is typically implemented by estimating a non-linear regression for the probability to participate in the intervention, followed by a (weighted) linear regression to estimate the impact of the intervention.

Table 1. Types of impact evaluation designs

Experimental studies	Observational studies	
	Selection on observables and unobservables	Selection on observables
Randomised assignment (incl. over-subscription)	Instrumental Variables (IV)	Covariate adjustment
Conditional randomised assignment / (raised) threshold randomisation	Regression discontinuity design (RDD)	Statistical matching
Randomised phase-in	Difference-in-Differences (DID)	
Randomised encouragement		

Source: Authors, adapted based on Kluve and Stöterau (2014[13])

4.3. Parameters estimated in CIEs

The impact evaluation literature borrows a lot from the methodology and terminology established in medical sciences. Hence, the policy, programme or event under study are commonly referred to as treatments and the group of participants is called the treatment group[13]. The estimated impact is therefore called the treatment effect. As social sciences often face more complex evaluation settings, the methodological toolbox has been further refined. These refinements have given rise to different versions of the treatment effect. These different treatment effects, explained below, not only derive from different underlying methods but also hold different policy implications. However, in social sciences, the allocation of the treatment is often not as unambiguous as in a clear laboratory experiment to test a new drug.

The key challenge is that the evaluator can often only determine (and observe) the assignment to an intervention but not also actual participation. In the case of experimental studies, it is often the case that evaluators can only assign individuals to the treatment or control group; but those assigned to the treatment group can still decide whether to participate or not. In addition, they may even hide their actual participation status after being assigned to participation. Also in the case of observational studies, the evaluators can often observe who was (externally) assigned to an intervention (e.g. target group of a policy) but not always who actually benefitted from it. In both cases the evaluator cannot influence and/or observe who took up the "offer" to benefit from an intervention. The impact evaluation refers to both cases as non-compliance. Non-compliance can also occur among the comparison group, in case individuals are able to self-select into the intervention.

If there is non-compliance, simple impact estimates can be misleading. To estimate the impact of an intervention generally requires to observe who actually benefitted from an intervention. The question is who the non-compliers are and what drives their behaviour. The key issue is that selection in or out of the

[13] In practice, an evaluation can assess the impact of more than one treatment in order to compare among them. In this case, the evaluation would need to define multiple treatment and comparison groups. For simplicity, this note refers only to the basic case of one treatment and comparison group.

intervention may be systematically related to the potential level of outcomes among compliers and non-compliers. For example, in an RCT of a skills training for jobseekers, only those who are the most motivated to find a job may decide to participate, out of all those who were (randomly) offered the program (the compliers). However, their potential chances to find employment irrespective of the training are likely also higher than those of less motivated non-compliers.

In the case that an intervention observes some (intended or unintended) non-compliance different measures of impact can be assessed. First, the impact of assigning someone to the programme, irrespective of whether they eventually participated. This is called the *Intention-to-Treat Effect* (ITT). Second, the impact of actual participation in the programme, called the *Average Treatment Effect on the Treated* (ATT). The ITT and the ATT measure different impacts, which are both important indicators for the effectiveness of a programme. Deciding which is more relevant to inform policy depends on purpose of the evaluation. Often, the ATT will be of greater interest for policy-makers as it represents a more immediate indication about the effectiveness of service providers to deliver services. In addition, the ATT is usually the basis to calculate the cost-effectiveness of a policy. The ITT, however, may represent a better basis to judge the expected benefits of up-scaling the pilot programme, as it takes into account that similar patterns of non-compliance can be expected. Moreover, estimating the true ATT usually requires more elaborate statistical methods and assumptions, if participants are indeed a selective subgroup of the random treatment group. In the case that there is full compliance[14] with treatment the ITT and ATT are the same. This is typically called the Average Treatment Effect (ATE) for the total population.

In the case of non-compliance, another option is to estimate Local Average Treatment Effect (LATE). The LATE is estimated when there is non-compliance in both the treatment group and in the comparison group. In this case, the LATE represents the impact of the intervention for the specific subgroup of the population that complies with their original assignment status. For example, in the case of a voluntary job-training programme, the LATE represents the impact of individuals in the treatment group who participate in the programme and who would have not participated had they been assigned to the comparison group. The LATE principal is an even more important concept in the framework of quasi-experimental studies where actual participation is not directly observable (see below). The results from estimating the LATE generally thus has to be interpreted carefully. The LATE estimation principles apply when there is non-compliance in either the treatment group, the comparison group, or both simultaneously. The ATT is simply a specific case of the LATE when there is non-compliance only in the treatment group.

4.4. The gold standard of policy impact evaluations: Experimental studies

The safest way to avoid selection bias is a randomised selection of the participant (treatment) and non-participant (control) group before the intervention starts. This evaluation design is commonly referred to as a Randomised Controlled Trial (RCT).[15] When the treatment group and control group are assigned at random (e.g. local public employment offices randomly selecting participants for an intervention) from the same eligible population, both groups will have the same characteristics before the intervention on average. The only difference after the intervention is that one group has been subjected to the intervention and the other has not. Consequently, in a well-designed and correctly implemented RCT, a simple comparison of average outcomes in the two groups can adequately resolve the attribution problem and yield accurate estimates of the impact of the intervention on the outcome of interest.

[14] That is, all individuals to whom the intervention has been offered actually participate, and none of the comparison group are able to participate in the intervention (cross-over).

[15] In the standard case, every eligible individual in the population will have the same probability to be selected for the intervention. In some cases, conditional (or stratified) randomisation is applied when it is desired that individuals with specific characteristics have a higher chance to receive the intervention.

In which cases to conduct pilots and experimental CIEs? How to randomise?

As discussed above, rigorous (experimental) evaluations are particularly useful when piloting policies prior to full implementation. The key goal of piloting is to provide an *ex-ante* indication of the benefits and challenges of a new programme or policy. But pilots can only provide a credible basis for evidence-based policy design if they are accompanied by an informative and valid evaluation. Experimental impact evaluations are one of the most informative methods to evaluate the effects and basis for cost-benefit analysis. But the beneficial connection between piloting and CIEs runs both ways: not only are rigorous CIEs particularly important to ascertain whether a pilot programme should be expanded, but pilots often also provide a very good basis to conduct experimental CIEs. In many cases, the policy maker may choose to provide the intervention to only a sub-group of the full target population. The reason could be either to restrict potential negative consequences or simply limited financial resources allocated to the pilot. Since there are less places in the program than eligible candidates, pilots may offer a natural chance to identify a control group in order to assess the impacts using a CIE. This is often more difficult, once the programme is implemented on a larger scale.

The assignment mechanism in a randomised controlled trial can be designed in different ways. The programme operational characteristics may determine how and when assignment can be allocated. The usual approach is the simple random selection within the full eligible group. However, in many setups it may not be feasible or desirable to assign programme participation purely at random among the eligible population. In practice, the programme operation rules often determine whether and how randomisation can be embedded in the intervention design.

- **Oversubscription design**: The oversubscription design refers to the classic random assignment that can be implemented when there are more eligible applicants/individuals than places available due to (e.g.) limited programme resources.[16] It means that randomisation takes place among all the eligible applicants/individuals. This design offers itself ideally to assess the impact of pilot programmes in order to decide for extension or scaling up.
- **Conditional/partial randomised design:** Sometimes, policymakers have three categories of potential beneficiaries in eligible population: i) some they absolutely want/need to provide the intervention; ii) some they don't want/need to treat, iii) some they would include if they had more resources. Randomisation could be conducted only within the third category.
- **Randomised phase-in (or pipeline) design:** This approach randomises the sequence in which participants are assigned to the intervention. In this case, the outcomes of later participants serve as the control group for earlier participants (outcomes at the same point in time). Typically, this is done by randomising roll-out across geographic regions, but one can also establish other criteria to determine earlier or later participants. In the phase-in design, contrary to other designs, all participants will eventually benefit from the intervention and become treated, although they do so at different times.
- **Randomised encouragement (promotion) design**: Instead of randomising participation, evaluations can randomly assign part of the eligible population to receive an encouragement to participate in the intervention. This incentive can entail an additional announcement or information about the programme or (monetary) compensation to partake in the program. This randomised encouragement serves as an external factor for the probability of receiving the treatment that is otherwise unrelated to the (potential) outcomes.

[16] Note that this is approach is distinguished from classic (basic) RCT, which just randomly assigns among all eligible units from a baseline survey (Ruiz and Love, 2012[10]), irrespective of actual oversubscription and thus (potentially) faces a number of the above mentioned constraints.

How to design pilots and RCTs? What to consider and what to avoid in the design? Ensuring internal and external validity.

Even though randomising treatment is in principle a comparatively straightforward approach, many aspects in the practical implementation can lead to misleading results. These challenges represent a threat to the internal validity of the evaluation. An evaluation is internally valid if it provides an accurate estimate of the counterfactual through a valid comparison group (Gertler et al., 2016[11]). Identifying the various risk that can lead to distorted ("biased") impact estimates often requires detailed understanding of the programme operational design and evaluation methods. Despite the clarity of a randomised evaluation methods, many practical factors need to be addressed in its implementation. They include resolving ethical issues, accounting for spill-overs to the control group, as well as for selective attrition, and ensuring heterogeneity in participation even if the programme is randomised.

To ensure that results provide a valid estimate of the true impact, it is important to maintain and assess the integrity of the evaluation design. Even if the evaluation achieved to design a random assignment on paper, implementation may be challenged in practice. Many of these challenges can be dealt with at the analysis stage, but the evaluator needs to collect the necessary data in order to be aware of the issues and able to address them. Some of the most common challenges include:

- **Low take-up or high dropout rates**: This often affects interventions that are poorly designed, not of interest, not easily accessible or poorly understood by the intended beneficiaries. If take-up is very low, the remaining sample size for the evaluation may not allow a valid analysis.
- **Spill-over effects**: Spill-overs happen when an intervention affects the outcome of non-participants, in either a positive or negative manner. Spill-overs can occur, for example, due to social interactions, intervention externalities, or equilibrium effects.[17] For the evaluation, spill-overs are a threat to the internal validity if they affect the outcomes of the control group sample. In this case, the outcomes of the control group do not adequately reflect the hypothetical outcomes in the treatment group had the intervention not taken place.
- **Unobserved non-compliance**: As discussed earlier, non-compliance refers to a discrepancy between random assignment and actual participation. If the actual treatment status is not known, the evaluation can only recover the ITT but not the ATT.
- **Attrition**: Attrition occurs when follow-up data cannot be collected for parts of the evaluation sample (e.g. changes in location, contact phone numbers). High attrition in itself can be an issue for the sample size of the evaluation. But attrition is only a threat to the internal validity if it is systematically related to assignment to treatment (called differential attrition). Plainly speaking, it "un-randomises" the initial assignment. Importantly, even if rates of attrition look the same, differential attrition may still occur if the characteristics of individuals leaving are related to treatment assignment. It is possible to test, whether differential attrition is present by using baseline data, if these are available.
- **Unintended behavioural effects:**
  - **Hawthorne effects**: The possibility that individuals alter their behaviour just because they are aware of being part of the treatment group in an evaluation (for example by increasing job search to demonstrate the value of a labour market programme).
  - **John-Henry effects**: The effect that individuals in the control group change their behaviour because they feel excluded from the programme.

[17] An explanation of these different types of spill-overs would go beyond this note. See Gertler (2016[11]) for a short explanation of each of them.

- **Anticipation effects**: Can be an issue for experimental methods when individuals in the treatment or control groups expect to receive the intervention and begin changing their behaviour before the programme actually reaches them.

To provide valid information for policy design, results from the evaluation also need to be generalizable. In most cases, evaluations are not conducted on the entire population of eligible individuals but rather on a sub-sample of them, for example because of limited programme or evaluation resources. The core necessary condition for an evaluation to provide informative, generalizable results is external validity. External validity means that the evaluation sample accurately represents the population of eligible individuals. One way to achieve this is to randomly sample the evaluation sample from the full population of eligible individuals. However, while this ensures that impacts identified in the evaluation sample can be extrapolated to the population, it does not yet imply that the results also hold in similar contexts or changes to the design (e.g. upscaling).

4.5. When RCTs are not possible: observational studies

When a randomised controlled trial cannot be conducted, non-experimental data based evaluation designs offer a viable alternative. In some cases, for instance when randomisation is not operationally feasible or ethical, there are alternative designs that an evaluator can choose. These approaches use statistical methods based on additional information to mimic randomisation (usually *ex-post*, after the programme has started). These methods hence compare individuals receiving the intervention with a non-random group of the eligible population (self-) selected to not receive the intervention. The empirical analysis aims to construct a comparison group that resembles the intervention group on relevant characteristics prior to the intervention. If this is successful, the programme impacts can be estimated with a reasonable degree of confidence.

The key difference among research designs for observational studies regards the assumption concerning selection into treatment. The relevant methods are typically classified by whether they account for selection on observable (reflected in the data available for the evaluator) or also unobservable factors (characteristics that are not observable for the evaluator).

Evaluation methods assuming selection to treatment on observable and unobservable characteristics

These evaluation methods seek to find an external ("exogenous") factor that changes the probability for some individual to be affected by an intervention. The key idea is that this external factor is unrelated to the (unobservable) characteristics of participants and non-participants that drive (self-) selection into the intervention. In many applications, this factor is a specific feature of the programme design, such as a predetermined cut-off for eligibility (e.g. by age). Another common application is to use a certain event as a factor that affects some groups of individuals or geographic areas more than other (e.g. a reform in one region).

Three types of research design are typically differentiated among the methods that account for selection on unobservables. These three methods differ in their underlying assumptions, data requirements and statistical techniques. Nevertheless, all these methods require that all factors which determine assignment to the programme are known and well understood.

The main methods assuming selection on observable and unobservable characteristics are the following:

- **Instrumental variables (IV):** In an IV regression, an indicator for the external factor (the "instrument") is simply included in the statistical model. Individuals with different values of the exogenous factor differ in their probability to participate, but are otherwise comparable. In

simplified terms, the inclusion of the exogenous variable "cleans" the (non-random) treatment variable from the non-random part. Thus only the variation in the probability to participate that is due to the exogenous factor remains in the model. Hence, it is possible to apply the RDD only in case it is possible to identify an exogenous factor (variable for which data exist also) that credibly defines the probability of treatment.

An example for MITES: Assuming that a jobseeker's geographical distance to the nearest training centre does not define training needs or labour market outcomes, but does define the probability to participate in a training, the distance to training centre could be used as an instrumental variable defining selection to treatment to evaluate labour market outcomes of jobseekers.

- **Regression discontinuity design (RDD**): The basic idea of RDD regression is to exploit some (non-random) cut-off point in the criteria that is important for selection into treatment. The design then simply compares individuals just below and above this cut-off. The underlying assumption is that those close to the cut-off are sufficiently similar to justify that ultimate selection was "as good as random". It is important that the individuals cannot precisely manipulate the variable defining the selection to treatment. Hence, it is possible to apply the RDD only in case a cut-off point in the selection to treatment exists, defined by some observable characteristics (i.e. the evaluator has the data that defines the selection to treatment).

  An example for MISSM: The amount of the new Minimum Income Scheme (MIS) is different for a single parent living with a child under 18 years of age and for a single parent living with a child above 18 years of age. The effect of the MIS on its expected outcomes could be evaluated for single parents around the cut-off point of the child's age. Nevertheless, the evaluation parameter will be LATE and cannot be extended to all single parents or all MIS recipients.

- **Difference-in-Differences (DID):** In essence, DID regressions simply compare outcomes of participants and non-participants before and after the intervention.[18] By estimating the difference in the differences in the outcome of interest before and after the intervention, all pre-existing differences are excluded from the analysis. The underlying assumption is that some unobserved factors for participation are present but that these factors are time invariant. In contrast to IV- and RDD-regression, the DID approach thus requires pre-intervention data on the outcome variable.[19] Hence, it is possible to apply the DID only in case data on outcome variable exists for both treatment and comparison group before as well as after the intervention.

  An example for MISSM: Some Autonomous Regions had benefits similar to MIS in place before MIS was introduced nationally. It is possible to study the effects of the effects of MIS additional to s regional scheme, by studying the desired outcomes of people who received the regional scheme before the introduction of MIS and continued to receive it, to the desired outcomes of people who received the regionals scheme before and were transferred to MIS. To conduct DID, the data on the outcomes of all benefit recipients need to be available both before and after the introduction of MIS.

These methods generally estimate the LATE – in particular Regression discontinuity design and Instrumental variable method. That is, the resulting impact estimates apply only to the subset of participants whose probability to benefit from the intervention was actually affected by this factor.

[18] Hence, DID combines the two naïve estimates of the counterfactual (before-and-after comparisons, and participant-non-participant comparisons) to produce a better estimate of the counterfactual.

[19] In the context of IV and RDD regressions, pre-intervention data is nonetheless highly desirable to test the main assumptions underlying this approach.

Evaluation methods assuming selection to treatment on observable characteristics only

These methods are typically employed when randomisation was not feasible *ex-ante* and no external source of variation can be found. The basic idea is that adjusting the measured outcomes of the participant and/or comparison group for differences in their characteristics (observed prior to the intervention) can remove all differences in the follow-up outcomes that are not related to the participation. The underlying assumption of these methods is thus that all factors that jointly determine selection into treatment and the potential outcomes of participants are (directly or indirectly[20]) accounted for by the statistical model. This assumption is usually referred to as "selection on observables". Hence, all factors that are thought to determine whether an individual decided to participate should be reflected in the available data. This can often be difficult, for example in cases were individuals can decide themselves and thus intrinsic motivation may play a key role. Thus, to be credible, these methods typically require a large amount of data about the treatment and comparison groups.

Methods that assume selection on observables can be distinguished into two broad statistical approaches, which differ in their assumptions on the data.

- **Covariate adjustment:** Covariate adjustment typically refers to all regression-based methods. The key idea is to adjust the means of observed outcomes in both treatment and comparison groups for the differences in other (observable) characteristics of individuals in both groups. This is achieved by including all available information one has of (pre-intervention) characteristics as variables into a statistical model (e.g. a linear or non-linear regression). By including these variables in the model next to an indicator for the observed treatment status, the estimate of the latter variable represents the average "net" effect of the treatment (e.g. "ceteris paribus" – all other factor equal).[21] Hence, these models can be used for impact evaluation only in case a rich dataset exists that is assumed to have all variables credibly defining the selection to treatment.

  An example for MITES: A simple evaluation could estimate an effect of a training programme on labour market outcomes by applying an ordinary least squares regression, where an employment status after 12 months from registering as an unemployed (dummy variable) is the dependent variable, and a dummy variable to indicate whether a person participated in a training programme during these 12 months as well as other personal characteristics are included as explanatory variables. The estimation of the covariate for training participation is the estimation of the effect of training on employment. Note however, that this estimation is likely biased as people participating and not participating in training can be different beyond the characteristics included in the regressions model.

- **Matching**: The intuition of matching-based methods is to generate a (sub-)sample from the remaining eligible non-participants that resemble the participant group as closely as possible. The most commonly known form is Propensity Score Matching. In its essence, propensity score matching involves two steps: In a first step, the individual probability of receiving the treatment is estimated given the observed characteristics (the "propensity score"). This is done by estimating a non-linear regression (e.g. a probit model). This model includes as the dependent variable whether an individual participated or not. The explanatory variables should cover all factors that are believed to determine selection into treatment. From this model, one can calculate the hypothetical probability that an individual participated given the observable

[20] The idea here is that not all underlying factors have to be observed and included in the model, as long as other variables in the model "sufficiently" control for the unobserved factors. Hence, the included variables have to be very closely correlated with the (unobserved) actual factor that drives participation. This is often difficult to justify.

[21] Simple regression models alone do not use a proper counterfactual and are not considered to be among the quasi-experimental designs. In most cases, this approach would not enable to receive unbiased evaluation results.

characteristics included in the model. In a second step, this predicted probability for each individual is used to estimate the treatment effect.

Propensity score matching methods differ how the predicted probability is then used in the second step. One approach is to select for each participant one or more individuals from the comparison group with a similar hypothetical probability (e.g. Nearest Neighbour Matching, Caliper or Radius, etc.). An alternative, is to use the predicted probability as weight in the following model to estimate the treatment effect (i.e. individuals with higher estimated probability to participate receive more weight). Examples of such reweighting algorithms are Inverse-Probability Weighting, Kernel matching or Entropy Balancing.

To be viable, matching methods thus need to ensure that the eventual treatment and comparison groups are "balanced" (i.e. sufficiently similar) in their observable characteristics after matching or reweighting. Furthermore, the evaluator needs to ensure that there is sufficient "overlap" in the predicted probabilities of treated and comparison groups – i.e. that one can actually find a credible comparison for each individual in the treatment group.

The most commonly cited technical guideline for matching by Caliendo and Kopeinig (2008[14]) therefore discusses six key steps to perform Propensity Score Matching: (1) Estimate the probability for each individual in treatment/comparison groups; (2) choose the matching algorithm that uses this probability; (3) check overlap between both groups ("common support"); (4) assess the matching quality (balance) and estimate the treatment effect; (6) perform sensitivity analysis on how the treatment effect differs when the methods or data is changed. Most statistical software packages include tools that combine most or all of these steps in a single command.

In summary, matching methods can be used for impact evaluation only in case a rich dataset exists that is assumed to have all variables credibly defining the selection to treatment. Compared to covariate adjustment, matching methods produce more credible evaluation results, as 1) they avoid any functional form restrictions, ii) they address the support problem that might arise in case some treatment group observations do not have similar comparison group observations, iii) they enable to handle better treatment effect heterogeneity (Lauringson, 2012[15]).

An example for MITES: Matching methods can be used to study the effects of training on labour market outcomes. The data needs are similar to the example described for covariate adjustment, but due to the different methodology, the evaluation results are more credible. Furthermore, matching and covariate adjustment can be combined by first conducting the matching and subsequently conducting covariate adjustment for matched observations.

An example for MISSM: A similar approach as described for MITES could be applied for MIS beneficiaries to evaluate the effects of applying labour market inclusion itineraries (i.e. matching the beneficiaries following the labour market inclusion itineraries to those who do not).

In some applications, various methods of both classes are combined to arrive at a more valid estimate of an intervention's impact. One example is to combine Matching and Difference-in-Differences, when simple matching cannot account for all unobserved characteristics that might explain why a group chooses to participate (and that might simultaneously affect outcomes). A special case of this combination is the Synthetic Control Method, which tries to offer a more systematic way to assign (probability) weights to the treatment control group. Similarly, Difference-in-Differences can improve the validity of Instrumental Variable regressions in some cases.

4.6. Technical tools (software) to apply micro-econometric methods

In most cases, conducting an impact evaluation requires knowledge of statistical software specific for analysis of micro-level data. While simple computation may be possible with standard spreadsheet software (e.g. Microsoft Excel), in many cases the data processing and empirical analysis will require software tailored to micro-econometric analysis. The most commonly applied statistical tools for evaluators today are Stata, SPSS, and R. These software packages differ in terms of their strengths, weaknesses, and handling of data, and choosing the right statistical software can make the work significantly easier.[22] It may be difficult and costly to switch halfway through a project, so the decision as to which software is the best fit should be made with care. It is important to take into account the technical abilities of the evaluators (both internal and external if contracting out the CIE) and the type of data that will be used (World Bank, 2020[16]). For example, conducting micro-econometric analyses on longitudinal data and larger datasets may require Stata, SPSS or R, while the data resulting from a neatly designed RCT could technically be analysed using Excel.

A software like Stata may facilitate conducting CIEs through the large number of built-in functions and user-written programmes. One advantage of using Stata, and also the reason of its popularity among researchers, is the large number of existing functions and programmes, which accommodate most micro-econometric methods previously described. Examples of such programmes are *rdhonest* to easily implement RDD, *ivregress/ivreg2* to implement 2SLS models with an instrumental variable, or *pscore* and *psmatch2* to construct a control group using propensity score matching. The existence of such pre-developed programmes simplify the use of Stata, which result in fairly low training needs to acquire the necessary level for collaborating with external researchers.

Upfront training costs for individuals to learn to use a free software such as R are arguably higher than for Stata, which offers a more coherent and streamlined user interface. Compared to Stata, R does not have as supportive user interface available allowing users to visualise the data, write and execute code as easily. There are also still somewhat less (online and free) training material and support forums for R than for Stata.

4.7. Skills needed to apply the micro-econometric methods

Conducting a full impact evaluation requires a broad range of skills and expertise of the programme design and context. The (internal or external) team responsible for evaluation should combine:

- Knowledge of the regional and institutional context
- Content knowledge (i.e. in the sector or thematic topic)
- Language proficiency
- Prior experience in design and leading evaluations
- Technical skills in evaluation design and data analysis
- Familiarity with administrative data sources or data collection process
- Reporting and communication skills
- Project management skills

[22] For a brief comparison see, e.g. (Ghosh, 2019[31])

5. Data requirements to conduct impact evaluations

5.1. Administrative data or survey data?

Monitoring and evaluating policies and programmes to assess their impacts requires information about outcomes of interest. Data are therefore the key component of any CIE, which warrants significant investments into data collection, processing and data management (OECD, 2018[18]). Different sources of data exist, each with its advantages and disadvantages. Figure 1 summarises the potential sources of relevant data, and specifies some of their principal characteristics.

Figure 1. Potential data sources to carry out CIEs

Survey data		Other data	
Experimental data *e.g. follow-up surveys*	**Observational data** *e.g. social surveys*	**Administrative data** *e.g. records from public agencies*	**Other types of Big Data** *e.g. social media, supermarket transactions*
• Data are collected to investigate a fixed hypothesis • Usually relatively small in size • Known sample / population • Usually not complex to use for research.	• Data specifically designed for research, may be used to address multiple research questions. • Data may be large • Known sample / population • Usually not very complex to use for research.	• Data are not collected for research purposes. • Data may be very large • Usually a known sample / population • Can be complex to use for research and require extensive data management to clean and organise the data • Multidimensional (i.e., may involve multiple fragments of data that have to be linked together).	• Data are not collected for research purposes. • Data may be very large • Sample / population unknown • Complex to use for research and requires advanced techniques to make it usable • Multidimensional (i.e., may involve multiple fragments of data that have to be linked together).

Source: (OECD, 2020[9]), based on Connelly et al. (2016[19]).

Administrative data are a powerful resource as they allow generating evidence with a high level of applicability for policymaking (Harron et al., 2017[20]). Some of the main advantages of administrative data over survey data are:

- **Cost-effectiveness.** Administrative data offer the possibility to evaluate policies using information that would be too costly to obtain otherwise. Administrative data do not impose major additional data collection cost and spare citizens from the burden of having to actively report the information. In addition, administrative data can be reused, while experimental data rarely can, as these are often very specifically tailored to a fixed hypothesis.
- **Greater population coverage.** Because of the cost involved in collecting data, experimental and observational data have smaller sample sizes than administrative data, which (ideally) cover the entire relevant population. This is necessary, as the advanced econometric techniques often used in CIEs would otherwise face the issue of too small number of observations (see Box 1).
- **Continuous recording of information.** The cost of collecting data also results in observational data being only collected at fixed periods, which does not make them suitable to analyse many policies and potentially longer term effects. Because administrative data are continuously recorded, it is possible to identify cohorts who experienced a particular policy change to study change over time, even if there was no survey data collection at the time (Connelly et al., 2016[19]).
- **Non-response, sample selection and non-random attrition.** The legal obligation to participate in administrative data programmes is a key advantage compared to the voluntary nature of responses to surveys, which limits the problem of non-response (UNECE, 2011[21]), as well as potential issues of sample selection and non-random attrition. Survey data may also suffer from other data quality issues, such as measurement error in responses due to misreporting and misunderstanding of questions. However, these issues often affect administrative data, too.

In many instances, administrative data may need to be complemented with other data sources. While administrative data offers an excellent population coverage, its information is limited to the aspects that are relevant for administrative/operational purposes. Depending on the objective of the intervention and of the CIE, evaluators may need to rely on other sources of existing data or may need to collect their own data. Legal and technical challenges may arise from combining administrative data with data from other sources, which requires a timely assessment of the data needs.

Working with administrative data requires a large initial investment in a good and agile data management system to make the data usable for research. To ensure that the wealth of data collected by MITES and MISSM can be used to evaluate their policies, the collected data must be of quality, and convertible to a format that can be easily analysed with a software suitable for microdata analysis. This is a comparative disadvantage of administrative data relative to survey data, which are usually already in the appropriate format for their analysis.

To ensure making the best use of the data, building a liaison between data owners and the analytical teams using the data is essential. An ongoing exchange between data owners and researchers also allows developing a system to give feedback on any quality issues of the data, which can be of value to the data owner, supporting future improvements. It may even allow influencing the data collection process, in often quite subtle ways, making them more amenable to CIEs (Statistics Canada, 2019[22]). **Working with administrative data requires addressing data protection concerns.** The use of administrative data may raise concerns about the privacy of the information in the public domain, particularly when the records are linked across different sources and data are stored for long periods.

5.2. Data needs to define the treatment group

The key requirement to conduct a CIE is to be able to identify participants to a policy or programme from the data. Evaluators need to access the registers recording participation in a given policy or programme. Participation records need to be designed in parallel to the design of a new policy, and need

to include at minimum a unique personal identifier, the start and end date of participation, and all possible information regarding the different activities and their intensity (European Commission, 2020[23]). The existing division of competences in Spain implies that it is likely that regional entities will record the data on participants to the policy to be evaluated. This is the case for most labour market policies, where Autonomous Communities record the information in a platform that automatically reaches MITES (SISPE system). MISSM collects the information on MIS participants centrally in a single database.

Records on participation in a policy should be integrated with data including characteristics on the eligibility to the programme. Personal characteristics that grant eligibility to a policy are very often key to conducting CIEs (e.g., when using RDD designs), but also to cross-check that participants to a policy meet the eligibility criteria. This can be done either through a unique register system, or by linking data across registers. In the case of the MIS of the MISSM, for example, this implies wealth and income information allowing to identify economically vulnerable individuals (and households), residence information, age, registration with SEPE and application and receipt of any other welfare benefit.

5.3. Data needs to define the comparison group

When there is a randomised assignment or the method to conduct the CIE assumes selection on observables and unobservables, the data needs to define a comparison group are generally lower. To construct a valid comparison group from non-participants, data are needed at least about programme participation, eligibility and target population. Additional information on comparison and treatment group characteristics is however desirable to verify that the treatment and comparison group are truly similar, and to apply matching in case they are not. In the case of the MIS, the target population are individuals residing in Spain for over one year. Since the MIS sets a maximum threshold of income and wealth for eligibility to the benefit, data on income and wealth on non-participants could be used to construct a control group of residents not eligible to the policy by a margin.

For research designs that assume selection on observables, e.g. covariate adjustment or matching, constructing a credible comparison group involves additional data needs. When using these methods, the evaluator requires a wealth of characteristics on programme participants and non-participants to perform the matching or covariate adjustment. These approaches, while commonly used, warrant a word of caution: eligible non-participants to a policy may be more similar in observable characteristics to participants (i.e., treated group) but there is a risk of bias from unobserved characteristics that caused their self-selection out of the policy. While the MIS may allow for an evaluation method assuming a selection on observables and unobservables, it is likely that the labour market policies of MITES have to be evaluated assuming selection on observables only. The target group of labour market policies of MITES are firms or unemployed, e.g. for training programmes. A comparison group can be constructed using matching methods on the pool of non-participant firms or jobseekers. To ensure getting the closest comparison group as possible, the evaluator will need a wealth of data, including demographic characteristics such as gender, age, education, residence, nationality, civil status, and in some cases, household level characteristics such as family composition, and characteristics of the family members. While most MITES registers record basic individual data, obtaining household data requires accessing more information from the Spanish Statistics Agency (INE).

Availability of longitudinal data allows constructing a better control group. When longitudinal data are available, it is possible to construct a comparison group by matching non-participants to policy participants on their pre-treatment outcomes in addition to time invariant characteristics. It also opens the possibility to implement a Difference-in-Differences methodology, on its own or in combination with Matching. Performing a Difference-in-Differences regression is likely to produce more credible estimates of the impact of a policy. While obtaining longitudinal survey data is often challenging, administrative data offers a clear advantage.

5.4. Data needs to define outcome variables: Key outcome variables for MITES and MISSM

The data needs to conduct a CIE depend on the policy or programme being evaluated and the research question. The objectives of the policy, and more broadly the mandate of the public institution, determine the research question and thus the data needs to construct the outcome variables of the CIE (European Commission, 2020[23]). Data availability determines the methodology that can be used to recover the causal effects of a policy. At the same time, the methodology used determines the baseline or pre-intervention data needed, and the subsample required for the evaluation.

The MISSM mandate, and in particular the mandate of the General Secretariat for Inclusion and Social Welfare Objectives and Policies in MISSM (SGOPIP), is to promote inclusive growth and reduce inequality. For example, the flagship new policy of the MISSM, the Minimum Income Scheme, should be evaluated on outcomes that are aligned with the mandate of the ministry, namely the impact on inclusiveness, as well as poverty and inequality reduction. More precisely, evaluating the impacts on:

- **Labour market inclusion**, by looking at employment, labour earnings and other job characteristics, as well as registration with the Public Employment Services (SEPE) and participation in ALMPs of population groups typically excluded from the labour market. This information is available in the SEPE register SISPE (Public Employment Information System)..
- **Social inclusion and wellbeing.** Social inclusion can be assessed by looking at outcomes measuring the (degree of) involuntary exclusion of individuals and groups from society's political, economic and societal processes, which prevents their full participation in the society (International Initiative for Impact Evaluation, 2014[24]). One such outcome is the school attendance and educational attainment of children of populations at risk of being socially excluded. Other outcomes can be measuring negative social attitudes and discriminating behaviours. Wellbeing could be addressed by measuring self-reported indicators, such as self-esteem and health outcomes. All these indicators are not available in administrative data, and thus would require collecting data through a tailored survey administered to programme participants and a control group.
- **Poverty reduction.** Assessing the impact of a policy on poverty is challenging, and the first step relies on making sure that some of the policy participants can be classified as poor. In this case, it is key to evaluate the effects by poverty level, and think about measuring poverty as an outcome (Goldstein, 2014[25]).
  - Possibly the most relevant outcome to measure poverty is **income**. Income is essential to measuring the incidence, depth and severity of poverty and inequality by allowing to construct income poverty (poverty headcount, poverty gap etc.) and income inequality outcomes (Gini coefficient, inequality ratios etc.). In Spain, income is recorded in the administrative data register "Income, tax and households information", owned and managed by the Tax Agency. Data on **wealth** is also relevant, particularly when it comes to measuring inequality.
  - While income is the most common way of measuring poverty, it comes with its challenges. Income-based poverty indicators do not usually include non-cash benefits, and require special care on how to treat taxes. A **consumption**-based poverty indicator may be less volatile, and account for savings usage, access to anti-poverty programmes and ownership of durable goods. It would allow identifying those who live above the income poverty line but spend the majority of their income on food or healthcare, unable to afford proper housing (Meyer and Sullivan, 2012[26]). For the same reasons, measuring inequality through changes in consumption across different centiles of the population may be a good alternative to income inequality measures. In the Spanish set-up, consumption data are not recorded in

administrative registers, and would thus require collecting the information through a survey[23] or engaging in a collaboration with a consumption data owner (e.g. commercial banks' transaction records are often used to study consumption patterns of their clients).

- Lastly, the impact of the Minimum Income Scheme could be measured against the **resilience** of programme participants to withstand shocks. By assessing their vulnerability or preparedness to face shocks, it is possible to evaluate whether the policy has the potential to have long lasting effects on poverty and inequality reduction. Outcomes can thus relate to the access to credit and insurance. An alternative approach to evaluate the impact of the policy on individuals' and households' resilience is to track the impacts in the long term. Again, this information can be obtained through collecting survey data or through collaborating with a big data owner (commercial banks, insurance companies etc.).

MITES labour market policies, as well as approaches and tools used by public employment services are generally implemented with the goal of increasing the employment rate of the clients of public employment services. These interventions should thus be evaluated based on their impact on labour market outcomes of the participants/beneficiaries. In particular, the impact on employment, labour earnings, job sustainability, career progression, occupational match and other job characteristics should be evaluated in the medium and long term. This longitudinal information is available in the SEPE register SISPE (Public Employment Information System). Since these data are in-house for the MITES, agreements for the internal evaluating team to access them should be straightforward. However, it is important to note for the planning of evaluations involving external researchers that accessing SISPE data requires an official request of the data, a sound legal basis to use the data and a contract for data exchange.

[23] For example, the Family Budget Survey (*Encuesta de Presupuestos Familiares*) by the Spanish National Statistics Institute collects data on household expenditures on consumption by consumption areas: https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736176806&menu=ultiDatos&idp=1254735976608.

6. A roadmap to apply the Impact Evaluation Framework

A rough outline of the key elements to consider when evaluating a specific policy or programme is provided below. This should not be seen as a linear process as many of these elements are interdependent or some of them may not be relevant in certain contexts. The below list should rather be regarded as a checklist of the main aspects, rather than a step-by-step roadmap to follow.

Phase 1: Understanding the needs for counterfactual impact evaluations

- **Choose whether/which programme or intervention to evaluate, involving all relevant stakeholders.** Provide a very initial rough assessment of the feasibility, the possible costs and the likely benefits of conducting an evaluation of the respective intervention.
- **Establish the evaluation purpose.** Assess whether the main purpose is to ensure accountability for stakeholders, learn about the programme results for future policy design or something else.
- **Decide the evaluation criteria.** Given the evaluation purpose, which are the main criteria to determine success of the intervention (e.g. according to OECD criteria, such as relevance, coherence or effectiveness).
- **Determine the evaluation approach.** Given the evaluation criteria, decide about the suitable evaluation approach (e.g. process evaluation, impact evaluations (CIE), cost-benefit evaluation).[24]

Phase 2: Planning counterfactual impact evaluations

- **Determine the outcomes of interest.** This step involves the decision what will be measured to determine success regarding the key evaluation questions (results indicators). These indicators should describe the intervention's overall objectives and results in operational and measurable terms.
- **Determine a Theory of Change.** The Theory of Change should provide a clear indication of the main underlying assumptions of the intervention design. The goal is to define questions that can be answered with the CIE. Evaluation questions typically depend on the chosen evaluation criteria and thus the purpose of the evaluation. For impact evaluations, the basic evaluation question is typically "What is the causal effect (or impact) of a programme on an outcome of interest?"
- **Assess ethical considerations.** Ensure that ethical standards will be fulfilled in the design and implementation of the evaluation and related data collection.
- **Establish an evaluation team.** The evaluation team should include a range of well-formed individuals with expertise in both the specific intervention, policy and evaluation methods.
- **Choose an impact evaluation method.** The key step in the design is to determine the (possibly multiple) evaluation methods that are feasible and provide credible results. The chosen method should be derived from the operational rules of the programme (e.g. determining whether an experimental study is feasible).

[24] This step may come to the conclusion that other evaluation approaches are more suitable in specific context. In line with the focus of this note, the following describes the steps for conducting a (counterfactual) impact evaluation.

- **Determine the eligible population and sampling frame.** To (randomly) select treatment and comparison groups, one needs to delineate clearly the full eligible population from which groups are drawn. Once the eligible group is defined, a next step is to establish a "sampling frame", which is a comprehensive list of all individuals in the eligible population.
- **Plan the data collection**
  - **Decide on using survey and/or administrative data.** If administrative data are available, these are often preferable since these data will typically involve lower effort and costs. However, if administrative data do not provide information on the chosen outcomes of interest, additional survey data may be required.
  - **Choose the sample size (including power calculations).** If survey data are collected, it may not be feasible to collect data covering the entire treatment and comparison groups. In this case, power calculations are a key step to determine *ex-ante* whether the foreseen impact evaluation design will be able to provide valid results given the sample size.
- **Assess potential risks to the evaluation.** Assess which factors may impede the internal validity of the evaluation design (e.g. selective survey attrition, spill-over effects, unintended behavioural effects…).

Phase 3: Implementing and managing impact evaluations

- **In case sampling is needed (e.g. entire datasets are too big): Draw the evaluation sample.** In case the evaluation is not conducted with the entire treatment and comparison group, the evaluation sub-sample should be (randomly) drawn from the full eligible list.
- **In case of RCTs: Conduct randomisation.** When an experimental evaluation is feasible, randomisation must be conducted prior to the intervention start. The randomisation may be administered in collaboration with programme managers and staff, but should be monitored by the evaluators.
- **Collect survey and/or administrative data.** This involves collecting either baseline and/or already follow-up data, depending on the chosen evaluation method and type of data.
- **Clean and process data.** This step is needed in most cases both for administrative and survey data. Typical steps involve cleaning the data from erroneous observations, combining various sources of data, and checking for the validity of reported information.
- **Analyse data.** The final step often involves three key parts, although more steps might be needed dependent on the analysis at hand: 1) test whether the main assumption of the chosen research design are fulfilled (e.g. baseline balance); 2) run the empirical analysis to estimate treatment effects; and 3) check how sensitive the estimated effects are to deviations from the underlying assumptions.
- **Draft the report(s) on the evaluation results.** The content and writing style of the report should consider the target audience (e.g. the description of methodology can be more detailed and technical when a report is targeting researchers, but focussing on key messages when targeting high-level policymakers).

Phase 4: Disseminating results, ensuring policy uptake and managing knowledge

- **Disseminate internally.** This involves discussing the evaluation results with key policymakers, programme managers and staff, as well as supporting policymakers to translate the evaluation results into action plans, redesigned policies and guidelines for policy implementers.
- **Disseminate externally.** The evaluation results should generally be made available to the civil society, irrespective of whether the initial evaluation purpose was for accountability or learning. This may include distilling the most important findings, publishing them and possibly involving designing an outreach strategy.

- **Learn for future evaluations.** This involves drawing lessons from the evaluation regarding internal analytical capacity, data availability and process.

References

Caliendo, M. and S. Kopeinig (2008), "Some practical guidance for the implementation of propensity score matching", *Journal of Economic Surveys*, Vol. 22/1, pp. 31-72, http://dx.doi.org/10.1111/j.1467-6419.2007.00527.x. [14]

Connelly, R. et al. (2016), "The role of administrative data in the big data revolution in social science research", *Social Science Research*, Vol. 59, pp. 1-12, http://dx.doi.org/10.1016/j.ssresearch.2016.04.015. [19]

Costantini, M. and I. Higginson (2007), "Experimental and quasiexperimental designs", *Research methods in palliative care*, pp. 87-91. [27]

European Commission (2020), *How to use administrative data for European Social Funds counterfactual impact evaluations: a step-by-step guide for managing authorities*, http://dx.doi.org/10.2767/721497. [23]

Gertler, P. et al. (2016), *Impact Evaluation in Practice, Second Edition*, The World Bank, http://dx.doi.org/10.1596/978-1-4648-0779-4. [11]

Gertler, P. et al. (2010), *Impact Evaluation in Practice*, The World Bank, http://dx.doi.org/10.1596/978-0-8213-8541-8. [8]

Ghosh, A. (2019), *What's the Best Statistical Software? A Comparison of R, Python, SAS, SPSS and STATA*, https://www.inwt-statistics.com/read-blog/comparison-of-r-python-sas-spss-and-stata.html. [31]

Goldstein, M. (2014), "Development Impact", *Do impact evaluations tell us anything about reducing poverty?*, https://blogs.worldbank.org/impactevaluations/do-impact-evaluations-tell-us-anything-about-reducing-poverty (accessed on 18 August 2020). [25]

Harron, K. et al. (2017), "Challenges in administrative data linkage for research", *Big Data & Society*, Vol. 4/2, p. 205395171774567, http://dx.doi.org/10.1177/2053951717745678. [20]

ILO (2013), *ILO policy guidelines for results-based evaluation: principles, rationale, planning and managing for evaluations*, International Labour Office, Evaluation Unit (EVAL). [29]

International Initiative for Impact Evaluation (2014), *Productive Safety Nets Gap Map: all populations*. [24]

IPA (2015), "Reproducible Research", *Best Practices for Data and Code Management*, https://www.poverty-action.org/sites/default/files/publications/IPA-Best-Practices-for-Data-and-Code-Management-Nov-2015.pdf. [17]

Kleinbaum, D., L. Kupper and H. Morgenstern (1982), *Epidemiologic research: principles and quantitative methods*, John Wiley & Sons. [28]

Kluve, J. and J. Stöterau (2014), *A Systematic Framework for Measuring Employment Impacts of Development Cooperation Interventions*, Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) GmbH. [13]

Lauringson, A. (2012), *The impact of the generosity of unemployment benefits on Estonian labour market outcomes in a period of crisis*, Tartu Ülikooli Kirjastus, https://dspace.ut.ee/bitstream/handle/10062/25656/lauringson_anne.pdf?sequence=1&isAllowed=y. [15]

Leeuw, F. and J. Vaessen (2009), *Address the attribution problem*, Network of Networks of Impact Evaluation, http://www.dmeforpeace.org/sites/default/files/Leeuw%20and%20Vaessen_Ch4.pdf. [6]

Loi, M. and M. Rodrigues (2012), *A note on the impact evaluation of public policies: the counterfactual analysis*, European Commission, https://publications.jrc.ec.europa.eu/repository/bitstream/JRC74778/lbna25519enn.pdf. [1]

Meyer, B. and J. Sullivan (2012), "Identifying the Disadvantaged: Official Poverty, Consumption Poverty, and the New Supplemental Poverty Measure", *Journal of Economic Perspectives*, Vol. 16/3, pp. 111-36, http://dx.doi.org/10.1257/jep.26.3.111. [26]

OECD (2020), *Better Criteria for Better Evaluation Revised Evaluation Criteria Definitions and Principles for Use*, OECD Publishing: Paris, http://www.oecd.org/dac/evaluation (accessed on 2 June 2020). [3]

OECD (2020), *Impact evaluation of labour market policies through the use of linked administrative data*. [9]

OECD (2018), *Good Jobs for All in a Changing World of Work: The OECD Jobs Strategy*, OECD Publishing: Paris, https://dx.doi.org/10.1787/9789264308817-en. [18]

OECD (2016), *Open Government: The Global Context and the Way Forward*, OECD Publishing: Paris, https://dx.doi.org/10.1787/9789264268104-en. [4]

OECD (2010), *Glossary of key terms in evaluation and results based management*, OECD Publishing: Paris, http://www.oecd.org/development/peer-reviews/2754804.pdf. [7]

OECD (2006), *Outline of Principles of Impact Evaluation*, OECD Publishing: 2006, http://www.oecd.org/dac/evaluation/dcdndep/37671602.pdf. [5]

OECD (2002), *Glossary of Key Terms in Evaluation and Results Based Management*, OECD Publishing: Paris, https://www.oecd.org/dac/evaluation/2754804.pdf. [2]

Ruiz, C. and I. Love (2012), *Impact assessment framework: SME Finance*, The World Bank. [10]

Statistics Canada (2019), *Statistics Canada Quality Guidelines*, Authority of the Minister responsible for Statistics Canada. [22]

Stern, E. (2015), *Impact Evaluation: A Guide for Commissioners and Managers*, Department for International Development, https://assets.publishing.service.gov.uk/media/57a0896de5274a31e000009c/60899_Impact_Evaluation_Guide_0515.pdf. [30]

UNECE (2011), *Using Administrative and Secondary Sources for Official Statistics- A Handbook of Principles and Practices*, https://www.unece.org/fileadmin/DAM/stats/publications/Using_Administrative_Sources_Final_for_web.pdf (accessed on 28 January 2020). [21]

Wooldridge, J. (2009), *Introductory econometrics : a modern approach*, South Western Cengage Learning, Mason OH. [12]

World Bank (2020), *Development Research in Practice: The DIME Analytics Data Handbook*, DIME analytics. [16]