



DPME Evaluation Guideline 2.2.13

Guideline on Impact Evaluation

Created March 2014
Updated: March 2024

Addressed to	Government departments who are undertaking evaluations (programme managers and M&E staff) as well as evaluators of government programmes and policies.
Purpose	The purpose of this Guideline is to provide technical guidance on planning, undertaking and managing an Impact Evaluation to evaluators and commissioners of evaluations.
Policy reference	This guideline should be read in conjunction with the National Evaluation Policy Framework, 2019 (available on the DPME website).
Contact person for this guideline	Evaluation Unit E-mail: evaluations@dpme.gov.za Tel: 012 312 0162

Contents

1. Introduction.....	2	8. Typical Impact Evaluation questions.....	8
2. Purpose and rationale for the Guideline.....	2	9. Methodological Approaches to Answer Policy Relevant Questions.....	9
3. Definition of Impact Evaluation.....	3	9.1. Selecting the evaluation approach.....	9
4. Impact Evaluation: Conceptual Clarity.....	4	9.2. Designing the baseline survey.....	9
5. Ideal conditions for conducting impact evaluations	6	10. Choosing the appropriate design and methods.....	11
6. Purpose of impact evaluations.....	6	10.1 Experimental and quasi-experimental methods.	11
7. Key elements of designing an impact evaluation – using a Decision Tree.....	7	10.2 Alternative and complimentary methods.....	13
7.1 Using a Decision Tree.....	7	11. Planning, Prioritizing, Implementing and Managing Impact Evaluation.....	14
7.2 Decision Tree for Selecting Evaluation Design to Deal with Selection Bias.....	8	11.1. Undertaking an Evaluability Assessment.....	14

11.1.1 How to undertake an evaluability assessment ...	14	11.4.1 Who undertakes the evaluation	16
11.1.2 Checklist for programme evaluability	14	11.4.2 Terms of reference.....	16
11.2 Prioritizing interventions for Impact Evaluation.	15	11.4.3 Data Sources	17
11.3 Planning and Managing.....	15	12. Peer Review	18
11.3.1 Relevance and timeliness	15	Annex 1: Decision Tree for Selecting Evaluation	
11.3.2 Legitimacy.....	16	Design to Deal with Selection Bias.....	19
11.3.3 Credibility of the evidence.....	16	Annex 2: Glossary	20
11.3.4 Trade-offs	16	Bibliography.....	21
11.4 Managing Impact Evaluation	16		

Acknowledgments

The Department of Planning Monitoring and Evaluation (DPME) in partnership with the Centre for Learning and Evaluation and Results for Anglophone Africa (CLEAR-AA), would like to express their sincere appreciation to following contributors who supported the revision process: Dr Candice Morkel, and Ms. Khumo Pule from CLEAR-AA, Dr Tendai Gwatidzo from University of Witwatersrand Johannesburg, Ms. Mandisa Magwaza, Ms. Refilwe Keikabile, Mrs. Kgaugelo Moshia-Molebatsi, Ms. Thokozile Molaiwa, and Ms. Keketso Moloto from DPME, Mr. Deo-Gracias Houndolo - International Initiative for Impact Evaluation, Mr. Diniko Setwaba – National School of Government, Dr Neissan Besharati and Ms. Pamela Nyika from Pan-African Research, and Mr. Khotso Tsotsotso from Old Mutual Foundation.

We further extend our gratitude to Professor Sarah Chapman from the University of Cape Town for peer reviewing the guideline.

Your collective insights have greatly enriched the content and quality of the guideline. This guideline is proof of the power of collaboration and shared knowledge.

1. Introduction

The guideline was revised with the aim of providing an overview for government staff undertaking and managing evaluations. The guideline is broad and can be applied in different contexts. It provides the definition of impact evaluation as adopted by the Department of Planning, Monitoring and Evaluation (DPME), conceptual clarification of the term impact evaluation acknowledging various connotation of the term, its purpose, how to decide whether or not an impact evaluation is required, key elements of designing an impact evaluation, typical impact evaluation questions, methodological approaches to answering policy relevant questions, as well as analytical methods for use in impact evaluations.

2. Purpose and rationale for the guideline

The purpose of the guideline is to address the need emerging from the revised National Evaluation Policy Framework (NEPF), which includes impact evaluations as an evaluation type that departments and other organs of state can undertake as part of their compendium of evaluation approaches and methods. It seeks to further provide guidance on how to plan, manage and implement impact evaluations.

The rationale of the guideline stem from the rising demand for impact evaluation amongst policy makers, particularly in the context of a shrinking fiscus, protracted socio-economic problems, the ever-growing complexity of wicked social problems such as the climate crisis, poverty, food insecurity and inequality. Policy and decision-makers as well as citizens

alike are calling for better evidence of the impact of the investments that government is making in addressing these complex problems, and the global acceleration towards the achievement of the Sustainable Development Goals (SDGs) has further intensified the need for evidence of impacts.

In the global context, inadequate resources are being allocated to conducting impact evaluations, and documentation has shown that many UN agencies, multilateral development banks and governments are not adequately funding studies that provide evidence on which interventions work under what conditions, the difference these interventions are making, and at what cost. The Centre for Global Development (2006: 3) has termed this the “evaluation gap”, and warn that the tolerance for this gap is waning. The increasing demand from governments around the world for better evidence around “what works” is underscored by the contestation around the “right” path to development, which makes the need for evaluation in general, and specifically impact evaluation critical in addressing the development challenge.

Lastly, the national evaluation system in South Africa has grown and is maturing, and the capacity for different types of evaluation (particularly those of a more technical nature) are in demand. In the case of Impact Evaluation, there is a perception that these types of evaluations are difficult to do, that there is a small window of opportunity to “get it right”, and that it is best conducted by specialists in causal attribution and experimental designs. Little attention has been paid to the varying definitions of impact evaluation in practice, and how the sector can and must navigate the contested terrain between the various definitions and approaches.



3. Definition of impact evaluation

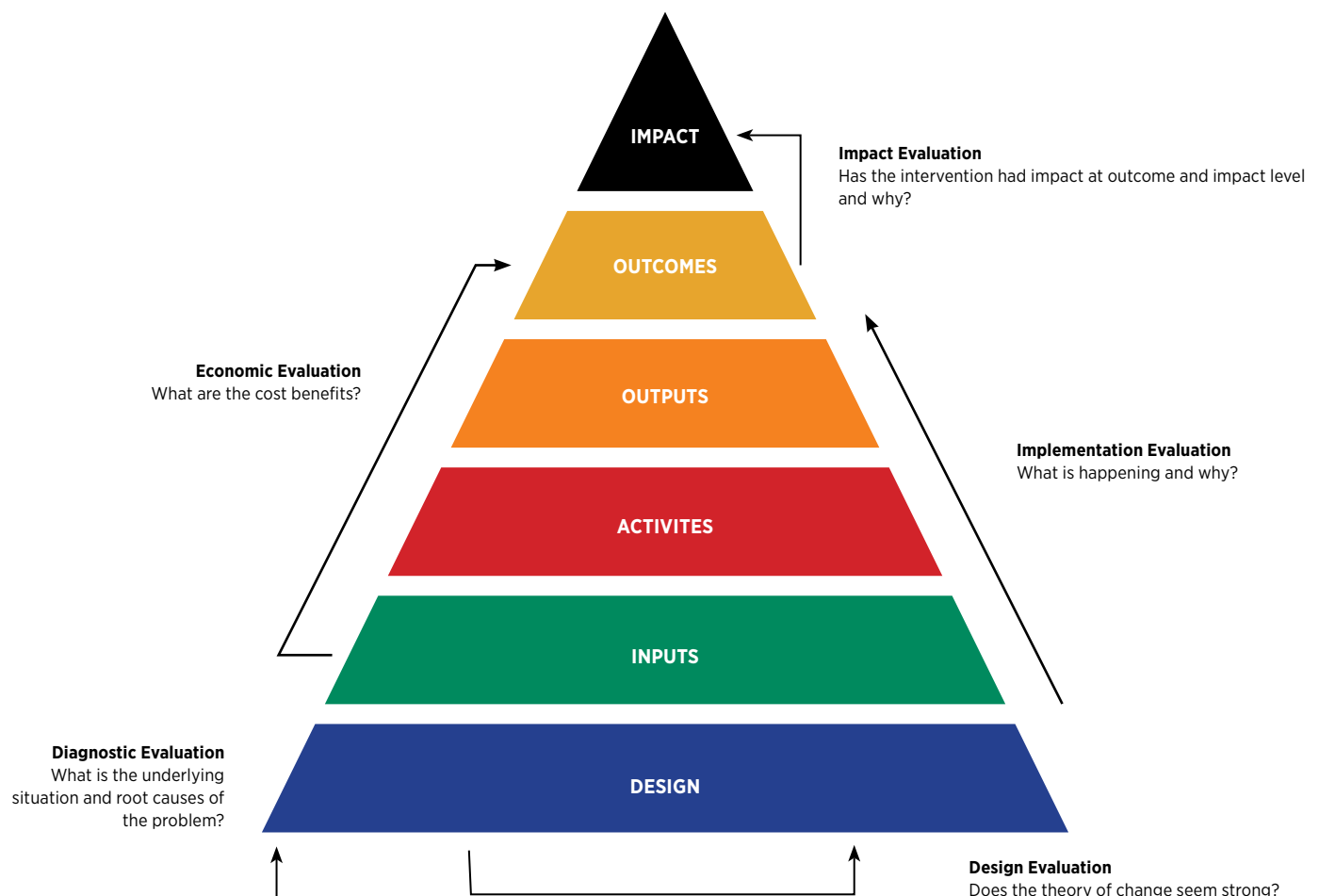
It is important to recognise that impact evaluation is a contested concept, with these contestations grounded in, firstly, different understandings and opinions of what constitutes “impact”, and secondly, in the choices made around designs, approaches and methods in determining impact. The DPME has adopted the definition of impact, which is used in international development evaluation. It defines impact as a change in the target population or social conditions that has been brought about by the programme/intervention (i.e. a change that would not have occurred had the programme or intervention not been implemented). In this definition of impact, an impact evaluation must establish the cause of the observed changes. Identifying the cause is known as ‘causal attribution’ or ‘causal inference’ (betterevaluation.org). Impact evaluations, in this case, inherently involves comparing the condition of targets that have experienced an intervention with an estimate of what their condition

would have been had they not experienced the intervention (Rossi et al, 2013). This is known as the counterfactual.

In section 4 below, a detailed discussion is provided to gain conceptual clarity around what is meant by impact evaluation in this guideline.

Impact evaluation can also be understood in terms of results-based management, and the hierarchy of achievement of results (for example the right combination of outcomes, taking assumptions into consideration and mitigating risks, will result in a certain impact, or medium/long-term effect). Figure 1 below, shows the results-based management pyramid which focuses on performance and achievement of outputs, outcomes and impacts. For example, achieving the outcomes of improved access to land and increased levels of participation in community decision-making might occur before, and contribute to, the intended final impact of improved health and well-being for women. The distinction between outcomes and impacts can be relative, and depends on the stated objectives of an intervention (betterevaluation.org).

Figure 1: Relationship of evaluations to results-based management



Source: National Evaluation Policy Framework, 2019

When a counterfactual is to be established, an impact evaluation needs information about impact(s) – but also information about inputs, activities, outputs and outcomes. It is therefore complemented by other types of evaluation found in the National Evaluation Policy Framework, 2019. This is a critical point, as an impact evaluation is not needed or advisable in all cases, and other types of evaluation may be more suitable for the questions that need to be answered, the purpose of the evaluation, or the stage of the intervention. Impact evaluations may also be conducted together with other types of evaluations, especially if there is a need to understand why and how impacts occurred, how positive impacts could be scaled up, and negative impacts avoided in future programming and policy making. Therefore, elements of an implementation evaluation design may form part of the design of a non-counterfactual impact evaluation.

In some cases, where the causal attribution between the output and outcome are direct and have already been empirically established, impact evaluations may just examine the extent to which outputs have been achieved. For example, the correct use of malaria bed nets in a programme directed at reducing the incidence of Malaria, or, the uptake of male circumcision in a programme directed at reducing the incidence of HIV/AIDs are all legitimate impact evaluations. In these cases, the intention is not to examine long-term effects.

A commonly asked question is when it is the “right” time to conduct an impact evaluation. For impact evaluations where causality is to be established, an impact evaluation should ideally be designed prior to implementation of the intervention the theory of change that underpins a programme or project can identify key points when it will be useful to collect data for an impact evaluation.

In the case of non-counterfactual impact evaluations, where the intention is not to ascertain attribution, but to observe longer-term, systemic effects, an impact evaluation needs to be conducted late enough for impacts, or longer-term, systemic or macro-outcomes to be evident.

This guideline firstly addresses the conceptual matters relating to impact evaluation and provides conceptual clarity to users of this guideline. The guideline also provides more technical guidance in the form of a Decision Tree that may be useful in deciding whether or not to conduct an impact evaluation, a typology of impact evaluation questions that may be posed, methodological approaches to answer policy relevant questions, and analytical methods for use in impact evaluations. These are not exhaustive lists but are illustrative and provide a foundation for further consideration in the process of designing an impact evaluation.

4. Impact evaluation: Conceptual clarity

The term “Impact Evaluation” has been contested and debated in the evaluation sector, owing to the various meanings ascribed to the term. The National Evaluation Policy Framework (NEPF, 2019) describes impact evaluation as measuring changes in outcomes (and the wellbeing of the target population) that are attributable to a specific intervention. From the perspective of measuring causal attribution, in its strictest definition, an impact evaluation:

“asks about the difference between what happened with the program and what would have happened without it”

(Center for Global Development, 2006: 12)

“is a study that attempts to measure the causal impact of a project, program, or policy on an outcome of interest to governments and other interested parties”

(Glewwe & Todd, 2022: 6).

“assesses the changes in the well-being of individuals that can be attributed to a particular project, program, or policy. This focus on attribution is the hallmark of impact evaluations. Correspondingly, the central challenge in carrying out effective impact evaluations is to identify the causal relationship between the project, program, or policy and the outcomes of interest”

(Gertler Martinez, Premand, Rawlings & Vermeersch, 2016: 4).

According to Gertler et. al. (2016: 8), the defining characteristic of impact evaluations is its focus on causal attribution, which determines the specific methodologies they will employ. They state that any method that is chosen to conduct an impact evaluation, must estimate the counterfactual – in other words, the outcome of the intervention on participants had they not been part of the intervention (Gertler et. al.). This must be done in order for evaluators to estimate the causal effect of a programme or intervention on its targeted outcomes (Gertler et. al.).

Impact evaluation provides information about the effects produced by an intervention. The intervention might be a small project, a large programme, a collection of activities, or a policy (betterevaluation.org). This definition acknowledges that impact evaluation:

- goes beyond describing or measuring impacts that have occurred seeking to understand the role of the intervention in producing these (causal attribution);
- can encompass a broad range of methods for causal attribution; and,
- includes examining unintended impacts.

Impact evaluations that are too narrowly focused on “average effects” in the pursuit of causal attribution can also be detrimental to the achievement of equitable development goals, which is a priority for the South African context of inequality. A programme that demonstrates an impact “on average” may actually increase equity gaps if the effects are not as effective for the most disadvantaged, yet are scaled up. The role of context (conceptually and empirically) is therefore critical in any impact evaluation (Betterevaluation.org).

However, many evidence consumers and end-users, including, for example, policy makers and citizens often use the term “impact” to mean the achievement of long-term, systemic or macro changes, or significant improvements that have occurred as a result of the implementation of policies and programmes. These changes may come about as a result of the implementation of a policy, programme or project, and may be observed at an outcome level (i.e., more immediately observed results) or impact level (i.e., longer-term effects). In this context, impact evaluations can therefore incorporate a multitude of relevant and appropriate methods and designs in order to answer these questions, they are not oblivious to the confluence of other variables that may contribute to changes, and are not intended to measure the counterfactual.

However, a cautionary note is issued here: the term impact evaluation has also sometimes erroneously been used to describe other types of evaluations, where some sort of



assessment of the achievement of outcomes has been undertaken. Glewwe & Todd (2022: 6) also suggest that a somewhat narrow definition of impact evaluation – i.e. randomised control trials (RCTs) – has also been erroneously promoted and popularised. A much wider interpretation has been accepted by many scholars and practitioners, including development partners, that includes numerous other types of methods and designs (Glewwe & Todd, 2022).

An argument that must be kept in mind is that “impact evaluations are just one type of evaluation” (Glewwe & Todd, 2022). Gertler et. al. (2016:7) agree that impact evaluation is one of many approaches that support evidence-based policy, including monitoring and other types of evaluation. It is commonly known that the design of an evaluation has been used as a “key marker” to identify whether an evaluation is rigorous and of high quality, and different designs are often placed in a hierarchical list, where (historically) randomised control trials (RCTs) had been placed at the pinnacle of what is considered the “best” evidence, and qualitative designs such as case studies near the bottom (Nutley & Powell, 2013: 11). Raimondo (2023: ix), in a publication aimed at contesting such ideas asserts that “several myths persist within research and evaluation circles about the power and limitations of evaluation designs that use cases (or case studies) as their primary empirical material (case-based evaluation designs)”. The paper dispels this myth by demonstrating how a World Bank evaluation of its support to carbon finance had been designed as a case-study, and dismisses false preconceptions “about the inferential, explanatory, and generalizability power of case-based evaluation designs” and “Many have moved away from the idea that any one methodology is the gold standard for all quality criteria for research” (Raimondo, 2023: ix, xi).

5. Ideal conditions for conducting impact evaluations

Impact evaluations can be undertaken at any point in the life cycle of a programme or intervention. For example, at policy formulation stage – an impact evaluation pilot can be undertaken to determine whether the proposed programme would actually have intended effects. When a new programme is at authorisation stage (started at a few sites) – an impact evaluation might be undertaken to show that the programme has expected effects before it is extended to broader coverage. During the implementation of interventions, an evaluation can be undertaken to enhance the effectiveness or to accommodate revised programme goals. When changes made are major, the modified programme might require an impact evaluation to be undertaken because it is virtually a new programme.

The simplest impact evaluations that are focused on causal attribution focus on simple cause-and-effect questions, for example whether a school health programme affects students' health and academic performance. A single outcome of interest, linked to a single intervention is the simplest and most basic expression of an impact evaluation of this nature. However, if there is more than one outcome of interest, the second (or subsequent) areas of interest would be estimated separately (Glewwe & Todd, 2022: 33). Critically, an impact evaluation should only be undertaken when its intended use can be clearly identified and when it is likely to be able to produce useful findings, considering the availability of resources and the timing of decisions about the intervention under investigation. It is recommended that an evaluability assessment be done first to assess these aspects (betterevaluation.org). In any circumstances when impact evaluations are conducted, there are pre-conditions that need to be met.

It is not always possible or feasible to conduct impact evaluations for all interventions, and they may be best conducted in cases where programmes are new or being considered for scaling up, but their effectiveness have not yet been established (Center for Global Development., 2006: 13). Undertaking impact evaluation would be useful in the following type of situations:

- Innovation schemes
- Pilot programmes which are due to be substantially scaled up
- Interventions for which there is sufficient or adequate evidence of impact in the given context
- A selection of other interventions across a portfolio on an occasional basis

When designing an impact evaluation, the following are key elements to consider:

- Deciding whether to proceed with the programme or intervention
- Identifying key evaluation questions
- The evaluation design should be embedded in the programme theory
- In the case of counterfactual impact evaluations, the comparison group must serve as the basis for credible counterfactual, addressing issues of selection bias (where the comparison group is affected by the intervention or similar intervention by another agency)
- Findings should be triangulated
- The evaluation must be well contextualised

6. Purpose of impact evaluations

Impact evaluation serves both objectives of evaluation: learning and accountability. A properly designed impact evaluation can answer the question of whether the program is working or not, and hence assist in decisions about scaling up. However, care must be taken about generalizing from a specific context. A well-designed impact evaluation can also answer questions about program design: which parts work and which parts don't, and so provide policy-relevant information for redesign and the design of future programs. We want to know why and how a program works, not just if it does.

In the NEPF, impact evaluation serves four different purposes:

- **Informing policy decisions** – the impact evaluation of large programmes, or the inclusion of impact evaluation data and findings in synthesis evaluations, can provide useful and convincing evidence to support decisions on public policy, including deciding which programmes will be funded in the future. Cost effectiveness and cost-utility comparisons can help compare different policy options, and as impact evaluations are measuring effectiveness, are usually possible.
- **Improving intervention design and implementation** – impact evaluations that can show impact but also explain how programs and projects work, and what is needed to make them work well, can inform and improve the design of future similar interventions. In this case they will combine with an implementation evaluation (see Guideline 2.2.12 on Implementation Evaluation).
- **Accountability** – Impact evaluation of government policies and programmes shows whether public funds are making a difference, and the extent to which the public interest has been effectively served. Even where

an impact evaluation finds that a programme or policy has not worked, the results can be used to improve the allocation of future resources, improving accountability.

- **Informing delivery** - Impact evaluations can be very useful in comparing different ways to deliver and implement a policy. In most development programmes, it is often delivery and implementation that requires information and informs the effectiveness of interventions rather than their efficacy.

Note that in all cases it should be combined with an implementation evaluation to understand if the intervention is working as designed or if there are important links in the causal pathway that are not being realised (uptake, behaviour change, adoption, delivery of inputs etc.).

7. Key elements of designing an impact evaluation – using a Decision Tree

7.1 Using a Decision Tree

A Decision Tree is a tool that can help users of this guideline to decide whether or not an impact evaluation featuring causal attribution and a counterfactual is needed, or whether a different evaluation design would best suit the evaluation purpose and questions. It may help answer the question: when is an impact evaluation to test causal attribution appropriate?

Table 1: Using a decision tree to design an Impact Evaluation

Question	If - Yes	If - No
1. Is this an evaluation of a specific programme or project?	Proceed to question 2	Consider other types of evaluations (see NEPF for guidance)
2. Is it your intention to conduct an experiment as part of part of your programme design?	Proceed to question 3	Consider other types of evaluations (see NEPF for guidance)
3. Was there a situation or problem analysis done before the programme was designed?	Proceed to question 4	Consider doing a diagnostic evaluation
4. Is there a clear Theory of Change? i.e.: • Is it embedded in Results-Based Management (RBM) principles (clearly delineated logic/causal logic of how inputs, activities and outputs contribute to, or lead to outcomes, mechanisms of change and assumptions) Clear articulation of the desired change (outcome statement/s)	Proceed to question 5	Consider doing a Theory of Change and design evaluation
5. Are there Specific, Measurable, Attributable, Realistic and Time-bound (SMART) indicators for the outcome(s) you are trying to achieve? i.e. • Do you have a defined and consistently implemented monitoring system to collect data? • Was the monitoring system part of the design from the beginning of the programme? • Have you done a baseline measurement of the indicators BEFORE the intervention?	Proceed to question 6	Consider first developing and implementing a Monitoring Framework and Plan that includes: • An indicator framework, with technical indicator descriptors • Data collection, analysis and reporting procedures • Baseline data collection
6. Are there evaluations, research, performance or monitoring reports that provide robust and trustworthy information on programme implementation? (i.e. whether it has been implemented according to plan, the effectiveness and efficiency of implementation; progress on activities and the delivery of outputs).	Proceed to question 7	Do an Implementation Evaluation
7. What was the ultimate outcome or impact you were hoping to have (on society, country, specific population)? Consider the maturity of programme: if it is too early, you may not be able to see results; if it is too late, you may start to lose results. Has enough time passed to see results? How long, normally (considering comparable interventions), does it take for this to show? Is it time to do an impact evaluation?	Proceed to question 8	Consult subject matter experts on: • Whether you need to wait some years to see these changes. • Whether other evaluation designs would be more suitable for your purposes.

8. Do you have meticulously collected programme performance data on the programme, the participants and on the outcomes of interest?	Proceed to question 9	Collect programme performance data
9. Do you have a counterfactual – a control group that has been observed from the beginning?	May consider doing a Randomised Control Trial or any other type of experimental design	Can consider: <ul style="list-style-type: none"> • Creating a counterfactual • Non-experimental design (econometric) • Quasi-experimental design • A before-after design Also see Sections 6 and 7 below for more options.

The basic steps in this decision tree are as follows:

1. If the evaluation is being designed ex-ante, is randomization possible? If the treatment group is chosen at random then a random sample drawn from the sample population is a valid comparison group, and will remain so provided contamination can be avoided. This approach does not mean that targeting is not possible. The random allocation may be to a subgroup of the total population, e.g., from the poorest districts.
2. If not, are all selection determinants observed? If they are, then there are a number of regression-based approaches which can remove the selection bias.
3. If the selection determinants are unobserved then if they are thought to be time invariant then using panel data will remove their influence, so a baseline is essential (or some means of substituting for a baseline).
4. If the study is ex post so a panel is not possible and selection is determined by unobservable, then some means of observing the supposed unobservable should be sought. If that is not the case, then a pipeline approach can be used if there are as yet untreated beneficiaries.
5. If none of the above are possible then the problem of selection bias cannot be addressed. Any impact evaluation will have to rely heavily on the program theory and triangulation to build an argument by plausible association.

7.2 Decision Tree for Selecting Evaluation Design to Deal with Selection Bias

Annex 1 shows decision Tree for Selecting Evaluation Design to Deal with Selection Bias

8. Typical impact evaluation questions

General evaluation answers many types of questions; however, impact evaluations are a particular type of evaluation that seeks to answer a specific cause-and-effect question: What is the impact (or causal effect) of a programme on an outcome of interest? (Gertler et. al., 2016: 8). Impact evaluations that seek to answer questions around the long-term, macro and systemic changes after the implementation of a policy, programme or project, may answer many other types of questions.

Imas and Rist (2009) suggest that evaluations can address three types of questions:

- **Descriptive questions.** These revolve around “what is”, and includes an examination of processes, conditions, organisational relationships etc.
- **Normative questions.** An evaluation asking normative questions is concerned with measuring “what is”, to what “should be”, and examines whether targets have been achieved, which can apply to all levels of the results chain (inputs, activities, and outputs).
- **Cause-and-effect questions.** This kind of evaluation is concerned with outcomes, and focuses on the difference an intervention has made on outcomes.

Examples of evaluation questions and sub-questions for impact evaluation include (adapted from Rogers, 2012):

What was the overall impact of the intervention?

- Did the intervention (programme, project or policy) work? Did the intervention produce the intended impacts in the short, medium and long term?
- Was the impact attributable to the policy/programme under review?
- For whom, in what ways and in what circumstances did the intervention work?
- How much did the intended beneficiaries benefit and to what extent did the impacts meet their needs?
- What unintended impacts (positive and negative) did the intervention produce?
- Much broader - is this the best intervention to achieve the desired outcome?

9. Methodological approaches to answer policy relevant questions

What is the nature of the impacts and their distribution?

- Are impacts likely to be sustainable and durable?
- Did these impacts reach all intended beneficiaries? If not, why not?

What other factors have influenced the intervention to achieve impact?

- How did the intervention work in conjunction with other interventions, programmes or services to achieve outcomes?
- What helped or hindered the intervention to achieve these impacts?

How did the intervention work to achieve (or not to achieve) impact?¹

- How did the intervention contribute to the intended impacts?
- What were the particular features of the intervention that made a difference?
- How do variations in implementation strategy affect impact? How have variations in the quality of implementation affect impact in different sites?
 - o To what extent are differences in impact explained by variations in implementation?
 - o Much broader - what is the best way to implement a given policy?



9.1. Selecting the evaluation approach

A major concern in selecting the evaluation approach is the way in which the problem of selection bias will be addressed. How this will be done depends on an understanding of how such biases may be generated, which requires a good understanding of how the beneficiaries are identified by the programme.

9.2. Designing the baseline survey

Ideally a baseline survey will be available so that double difference estimates can be made. Important principles in designing the survey are:

- Conduct the baseline survey as early as possible.
- The survey design must be based on the evaluation design which is, in turn, based on the program theory. Data must be collected across the results chain, not just on outcomes.
- The comparison group sample must be of adequate size, and subject to the same, or virtually the same, questionnaire. Whilst some intervention-specific questions may not be appropriate, similar questions of a more general nature can help test for contagion.
- Multiple instruments (e.g., household and facility level) are usually desirable, and must be coded in such a way that they can be linked.
- Survey design takes time. Allow six months from beginning design to going to the field, though 3-4 months can be possible. Test, test and re-test the instruments. Run planned tabulations and analyses with dummy data or the data from the pilot. Once data are collected one to two months are required for data entry and cleaning
- Avoid changes in survey design between rounds. Ideally the same team will conduct all rounds of the survey.

Options when there is no baseline

Evaluations are often conducted ex post, at times without having proper baseline established. Under these circumstances the following options can be considered:

¹ This point covers questions related to implementation evaluation, when linked with an impact evaluation.

1. If treatment and comparison groups are drawn from the same population and some means is found to address selection bias (which will have to be quasi-experimental, since randomization is ruled out unless the treatment had been randomized, but if the program designers had thought of that they will have thought of a baseline also), then a single difference estimate is in principle valid.
2. Find another data set to serve as a baseline. If there was a baseline survey but with a poor or absent comparison group, then a national survey might be used to create a comparison group using propensity score matching.
3. Field a survey using recall on the variables of interest. Many commentators are critical of relying on recall. But all survey questions are recall, so it is a question of degree. The evaluator needs to use his or her judgment as to what it is reasonable to expect a respondent to remember. It is reasonable to expect people to recall major life changes, introduction of new farming methods or crops, acquisition of large assets and so on, but not the exact amounts and prices of transactions. When people do recall there may be telescoping (thinking things were more recent than they were), so it is useful to refer to some widely known event as a time benchmark for recall questions.
4. If all the above fail, then the study should build a strong analysis of the causal chain (program theory). Often a relatively descriptive analysis can identify breaks in the chain and so very plausibly argue there was low impact.
5. Triangulation. This involves drawing on a variety of methods, data sources and approaches in order to address potential limitation of using any single source or approach.

Different methods are needed for the different elements of an impact evaluation:

- (1) Clarifying objectives and values;
- (2) Developing a theory of change;
- (3) Answering descriptive questions;
- (4) Answering causal questions; and
- (5) Summarising evidence into an overall judgement.

The first of these needs to be part of evaluability assessment work in initiating an impact evaluation. Having decided on the way forward, Table 2 summarises other questions and the key methods of impact evaluation.

Table 2: Questions and methods for different impact questions

Purposes	Common impact evaluation questions	Common evaluation methods and approaches
Element 1: Clarifying objectives and values	What are desirable impacts and what are negative impacts? What is a desirable distribution of benefits? What are an appropriate (set of) indicators that can help measure these?	<ul style="list-style-type: none"> • Literature review • Desk review • Appreciative Inquiry • Community surveys • Participatory tools with stakeholders • Narrative
Element 2: Developing a theory of change	What is the theory of change underlying the intervention	<ul style="list-style-type: none"> • Theory of Change • Log frame, a results chain or an outcomes hierarchy.
	How is the theory of change working in practice?	<ul style="list-style-type: none"> • Outcome Mapping • Factual analysis • Implementation/process evaluation • Methods below
Element 3: Answering descriptive questions	What has implementation been like (what activities have been undertaken and what has been the quality of implementation?) What agencies, people and mechanisms have been involved in the implementation (or absent in the case of implementation failure) What changes have occurred (and for whom?) What has been the context in which the programme has been implemented?	<ul style="list-style-type: none"> • Re-analysis of existing statistical data • Surveys, administrative data, census data • Observation • Interviews/group interviews/focus groups • Participatory tools • Monitoring data • Process evaluations • Most significant change
Element 4: Answering causal questions	How far has the intervention caused the impacts, contributed to causing the impacts, or have the impacts in fact been caused by other factors? How much of the impact can be attributed to the intervention?	<ul style="list-style-type: none"> • Counterfactual methods • Randomised control trials • Comparison group analysis • Logically created or expert constructed counterfactuals • Identifying and ruling out alternative explanations
Element 5: Summarising evidence into an overall judgement	What is the overall judgement to be drawn from the above data?	<ul style="list-style-type: none"> • Numerical scoring • Rubrics • Cost-effectiveness and cost-utility studies • Consensus consultation/experts' panels

10. Choosing the appropriate design and methods

There has been widespread recognition of the need to increase the range of methods that are used for evaluation, including causal analysis (Raimondo, 2023; Stern, 2012), and few of these are provided below as a “menu” from which to select an appropriate design. However, the guidelines for selecting an appropriate design provided below is not exhaustive, and design and methods specialists should be consulted in cases where the guidance below does not provide you with the specific evaluation design that is required to answer your evaluation question(s). One of

the most important points being made in this guideline is to dispel the misconception that “causal claims can be built only on approaches involving analysis of large numbers of observations using counterfactual thinking” (Raimondo, 2023).

10.1 Experimental and quasi-experimental methods

The (experimental and quasi-experimental) methods outlined below are useful for generating an impact estimate; the size and significance of change brought about by an intervention. These methods are less able to answer questions about how and why impacts occurred, who was affected, how context played an influence and the extent to which impacts are generalizable.

Table 3: Experimental and quasi-experimental design methods

DESIGN	CONDITIONS THAT BEST SUIT THIS DESIGN
<p>Randomised Control Trials Randomised Control Trials (RCTs), also known as experimental designs, involve providing a robust comparison between one or more groups receiving an intervention (treatment group) and a group that does not receive the same intervention (control group) through randomly assigning participants to each group. This ensures there are no observable or unobservable differences (or bias) between the treatment and control, meaning that any differences in measured outcomes between the two groups can be reliably attributed to the intervention, not an unrelated factor.</p> <p>RCTs have been mostly used in medical science particularly clinical trials. In evaluation, RCTs are used to measure impact, where:</p> <ul style="list-style-type: none"> Reasonable sample sizes can be constructed to allow for tests to be carried out on data which have sufficient statistical power. The randomisation for the RCT can be feasibly and practically integrated into the intervention design before it is implemented. It can be confidently assumed that the intervention has no impact on the control group <p>Restricting the intervention is appropriate and does not cause undue ethical risks.</p>	<ul style="list-style-type: none"> Allows robust comparison between groups, minimising bias in sample selection When implemented correctly, considered to produce robust estimates of impact The assumption of no impact on the control group may not be plausible. ‘Blinding’ (where participants, those administering the intervention and researchers do not know who is in the treatment or control groups) is rarely feasible in social interventions. Best used where the mechanisms by which the intervention is expected to work are well understood: this is often not the case. Best used when there is little variation in the execution of the intervention: it requires rigorous and uniform execution. It can therefore lack generalisability.
<p>Propensity Score Matching Propensity Score Matching (PSM) is a statistical technique that enables evaluators to construct a counterfactual group to estimate the impact of an intervention. This is achieved by matching treatment observations to one or more control observations based on their probability of being treated (or their propensity score). This is calculated using observable characteristics that determine the likelihood of participation and varies between 0 and 1 (where 1 is 100% likely to be treated). By comparing the outcomes of interest between the two matched groups an impact estimate can be calculated.</p> <p>PSM can be used when RCTs are either not feasible or not desirable. In order to estimate a robust counterfactual PSM requires:</p> <ul style="list-style-type: none"> A varied dataset available for matching made up of pre-intervention data to estimate the propensity score (as the treatment may affect post-treatment characteristics) Recipient and non-recipient groups should have a number of group members with similar scores (called, ‘presence of common support’) The assumption that assignment to treatment is only dependent on observable characteristics (known as ‘confoundedness’ or ‘Conditional Independence Assumption’). 	<ul style="list-style-type: none"> Allows an estimate of impact where Randomised Control Trials (RCTs) are not appropriate The estimated impact is the average effect of the treatment on all those treated, rather than a marginal impact on a small subset of the treated group Where rich data on factors affecting participation and outcomes is available, it is possible to use all of this with relatively few assumptions about the precise nature of these effects. As matching is only based on observable characteristics, where treatment and outcomes are affected by unobservable, impact estimates will be biased, and it cannot be determined analytically that there are no such unobservable factors. The sensitivity of results to unobserved characteristics can be explored through sensitivity analysis and can be mitigated somewhat by the addition of a difference-in-difference to the evaluation. As a result, rich data on both treated and untreated individuals are needed, preferably from the same source. If different sources are used the data must be directly comparable. As matching can only be done on pre-intervention characteristics, these data either need to be time-invariant (e.g. gender, year of birth) or collected beforehand.

DESIGN	CONDITIONS THAT BEST SUIT THIS DESIGN
<p>Interrupted Time Series Analysis (ITSA) It is a quasi-experimental method to establish the causal effect of an intervention. ITSA uses time-series data to test whether there is a change in the trend of outcomes following the introduction of an intervention. ITSA is particularly useful when an intervention is implemented at population level (such as estimating the effect of a new law) and when there is a clear time point of introduction.</p> <p>ITSA does not require a control group. Without a control group, impacts are estimated by assuming that trends would continue in the absence of the intervention. The method therefore relies on the absence of other interventions or short-term time effects that might influence trends around the time of the intervention. If this is not plausible, those changes can potentially be estimated by reference to a control group which has historically followed similar trends, which is not subject to the intervention, but which is subject to the same external influences.</p> <p>ITSA requires time series data from before and after the intervention, and is ideally used with administrative data. A data series which is too short can impact the power of statistical tests and resulting estimates should be treated with caution.</p>	<ul style="list-style-type: none"> • ITSA produces internally valid estimates of intervention effects even in the absence of randomisation, assuming confounding factors are stable over time (i.e. no other interventions are introduced at the same time as the intervention that would affect outcomes, and relevant population parameters remains stable. ITSA can be implemented retrospectively using administrative data • ITSA works best when there is a clear intervention time point, although gradual or delayed intervention introduction can be accounted for. • ITSA requires sufficient time-series data to take account of seasonality, autocorrelation and non-stationarity
<p>Instrumental variables (IV) Instrumental Variable (IV) regression is a method of estimating impact that makes use of a different variable (the instrument) to predict treatment in an econometric analysis. An IV is a factor which influences participation in the treatment, but which otherwise has no impact on the outcome. Providing an instrument is found which meets these conditions, an unbiased estimate of the impact of the treatment can be derived.</p> <p>IV may be appropriate to use when:</p> <ul style="list-style-type: none"> • Interventions may have been placed in a biased way which would also effect outcomes (e.g. in areas with higher rates of deprivation) • Individuals may self-select suggesting they have characteristics that make them more likely to be treated, or that also affect outcomes (e.g. prior experience of a similar project) <p>There is time-varying selection bias, that is when individuals change their likelihood of treatment over time.</p>	<ul style="list-style-type: none"> • Useful in instances where other quasi- and experimental methods are not possible • Does not require assumptions about there being no other sources of selection bias. • Finding a valid instrument is difficult, and usually cannot be planned in advance. If the instrument is only weakly correlated with treatment, great care is needed to derive valid impact estimates. • The derived impact estimate is a Local Average Treatment Effect – the impact on those who are on the margins of participation.
<p>Synthetic Control Methods Synthetic control is a quantitative method which uses historical data to construct a ‘synthetic clone’ of a group receiving a particular intervention. Divergence between the treatment and its synthetic clone provide the impact estimate.</p> <p>The synthetic control method is often used at the macro-level for policy evaluation and is particularly appropriate when there are a small number of treated observations. A relatively common application is where the units of treatment are areas. The method requires a pool of potential comparable observations from which to draw a weighted average that approximates the treatment observation e.g. counties, villages. This weighted average is calculated using historical data and then continued through the time-series after implementation to form the ‘synthetic clone’.</p>	<ul style="list-style-type: none"> • The key advantage of this method is that it can create a relevant and highly visual point of comparison where no suitable comparators exist. • It may be particularly suitable for analysing the effects of policy interventions targeting specific local economic outcomes and other areas where large volumes of secondary data is already available. • The analysis is only viable where it is possible to establish a historical relationship between the behaviour of the treatment and control groups.
<p>Difference in Difference Impact is measured by studying the outcome of interest before and after the intervention for two groups; one of which was subject to the intervention and the other not. First, the trend lines for the outcome of interest for the two groups are compared for the pre-intervention period. Where these trend lines move in parallel over time, a counterfactual trend can be estimated for the treated group (group A), which is then used to estimate the impact of the intervention.</p>	<ul style="list-style-type: none"> • Method is intuitively simple, and easy to explain. • Relies on the assumption that the outcome variable for both groups would continue to move in parallel if the intervention had not occurred. • The quality of this method is strongly tied to the quality of the data used with a substantial amount of data often being needed. • As with experimental designs, sufficient sample size is required.

Adopted from: HM Treasury (2020) Central Government Guidance on Evaluation, Annexure A

10.2 Alternative and complimentary methods

The below are theory-based methods, which can be used for impact evaluation to address questions about whether

the intervention caused an impact, how and why it occurred, how context may have influenced outcomes and help understand to what extent results are generalizable. They allow attribution of causality, but none gives precise estimates of effect sizes.

Table 4: Theory based design methods

DESIGN	CONDITIONS THAT BEST SUIT THIS DESIGN
<p>Qualitative Comparative Analysis (QCA) It is a method used to compare different aspects of an intervention and contextual factors to understand the different characteristics or combinations of characteristics which are associated with outcomes. It enables systematic comparison based on qualitative knowledge. Rather than examining the factors causing a specific outcome in depth as in a single case study, QCA focuses on identifying a variety of patterns. This allows for both complex causation (combinations of factors) and 'equifinality' (multiple causes of an outcome) to be accounted for.</p> <p>It is useful when the context within which an intervention is implemented is likely to influence its impact. It can identify which factors are necessary for the success or failure of an intervention and to understand why an intervention has worked in some contexts (such as areas) but not others and also to compare the efficacy of smaller interventions within a wider programme.</p>	<ul style="list-style-type: none"> • A pragmatic method that can identify groups of causal factors that can reasonably be used in post hoc evaluation. • QCA works best when data on all the cases of interest are available and the number of cases is neither too small nor too large, around ten to fifty cases. • It can be used with larger numbers of cases however, depth of understanding will be necessarily reduced. • It may be difficult to determine which cases represent more 'success' or 'failure' than others
<p>Process Tracing It is a structured method to developing and assessing theories about how a particular outcome arose. It examines a single case of change and tests whether a hypothesised causal mechanism, proposed by a theory of change, explains the outcome. This allows for single cases to be examined where there is no counterfactual and multiple cases for comparison are unavailable. Process tracing can be used to test the contribution of an intervention to an impact. A hypothesised causal mechanism, or several, is identified using a theory of change.</p>	<ul style="list-style-type: none"> • Process tracing is a practical method for understanding and testing causal hypothesising 'real world' situations that can be used in ex-post evaluation of a single case. • This method must be used with rigour to prevent inferential errors; alternative explanations must be carefully considered. Equifinality should also be considered (i.e. the support of one causal mechanism may not preclude others).
<p>Contribution Analysis Contribution analysis is a method which is used to understand the likelihood of whether the intervention has contributed to an outcome observed, or not commonly known as contribution claim. Through a step-by-step process, it explores how the contribution would have come about using a broad range of evidence to test this. Contribution analysis can make use of a broad range of evidence types and can be used for all types of interventions no matter how complex the theory of change is. It can be used where it may not be possible to establish an experimental design testing cause and effect.</p>	<ul style="list-style-type: none"> • Useful where there is limited scope or opportunity to affect roll out of a programme (to allow for experimental methods) • Able to confirm or revise a theory of change. • The quality of the eventual analysis and contribution claim is dependent on the quality of the thinking about the attribution problem and theory of change. • Contribution Analysis does not provide definitive proof that the intervention has had acausal effect but rather an evidenced logical line of reasoning which gives some level of confidence of an intervention's contribution. • Works on average effects, therefore, should not be used if there is a large degree of variance about how a programme has been implemented or an expectation of different outcomes for different groups.
<p>Contribution Tracing It is a rigorous mixed qual-quant participatory method to establish the validity of contribution claims in evaluation, with explicit criteria to guide evaluators in data collection and measuring confidence in findings. Contribution Tracing (CT) is inspired by both the principles of Process Tracing and Bayesian updating (probability).</p> <p>It gathers evidence which supports (or is against) a contribution claim. Evidence is analysed using mathematical formulae (Bayesian updating) to put a numerical value on the level of confidence in a particular claim. It is a participatory method which involves consultation with all relevant stakeholders through a series of steps i.e., making the claim developing a Theory of Change holding a contribution 'trial' with all the stakeholders to establish what would prove or disprove the claim identification of alternative causes application of Bayesian confidence updating ('put a number on it').</p> <p>Some steps can be taken in parallel e.g., steps 1 and 2, and 2 and 4</p>	<ul style="list-style-type: none"> • Points to what evidence to look for and what it means in relation to the claim. It only uses evidence with the 'highest probative value' i.e. evidence with the power to increase or decrease confidence in a specific claim, so time is not wasted asking other questions. • Specificity of the contribution claim increases the conceptual precision, clarity and quality of theories of change. • Minimizes confirmation bias by using 'critical friends' during the contribution testing phase, who represent other plausible explanations of the observable change • Participatory and collaborative • Not so useful in answering how a programme compares with other programmes. • Schedule of undertaking needs to be right - the intervention needs to have been going for long enough for the 'traces' to be visible • Must spend equal time and resources on exploring other potential causes to ensure all views appropriately considered.

Adopted from: HM Treasury (2020) Central Government Guidance on Evaluation, Annexure A

11. Planning, prioritizing, implementing and managing impact evaluation

This section outlines critical areas to consider when planning, prioritising, implementing and managing an impact evaluation. It highlights challenges that may be encountered in relation to impact evaluations.

11.1. Undertaking an Evaluability Assessment

According to the OECD-DAC, evaluability is “the extent to which an activity or project can be evaluated in a reliable and credible fashion”. The evaluability assessment is defined as a systematic process that helps to identify whether a programme is in a condition to be evaluated, and whether an evaluation is justified, feasible and likely to provide useful information. Its purpose is not only to conclude if the evaluation is to be undertaken or not, but also to prepare the programme to generate all the

necessary conditions to be evaluated (United Nations Development Fund for Women (2009).

An evaluability assessment is usually undertaken at the beginning of a programme or project to determine whether it is feasible and appropriate to evaluate the effectiveness of a programme or project. Provided there is intent to evaluate an intervention, assessing its evaluability can usually be done for a small cost of the total evaluation budget. This is particularly relevant when done in relation to an impact assessment, and can prevent wasting valuable time and resources on a premature or inappropriately designed evaluation.

11.1.1 How to undertake an evaluability assessment

An Evaluability assessment can be undertaken through qualitative data collection methods such as desk reviews, secondary data analysis, and interviews with key stakeholders. A checklist serves an important purpose in this regard. It identifies specific points that can be discussed with relevant stakeholders to determine evaluability as well as what needs to be in place to prepare for an evaluation.

Steps for undertaking an evaluability assessment:



11.1.2 Checklist for programme evaluability

Evaluability Parameters	Key Questions	Yes	No
Programme Design	Does the programme clearly define the problem that it aims to change?		
	Has the beneficiaries of the programme been determined?		
	Does the programme have a clear theory of change/logic model?		
	Is the results framework of the programme coherently articulated? Do the outputs, outcomes and goal follow results chain logic?		
	Are the objectives clear, measurable and realistic?		
	Do proposed programme activities lead to goals and objectives		
	Does the programme have capacity to provide data for the evaluation		

Evalability Parameters	Key Questions	Yes	No
Availability of information	Does the programme have capacity to provide data for evaluation?		
	Does the programme have SMART indicators on key areas of intervention?		
	Does the baseline information exist?		
	Does the programme have a monitoring system to gather and systematize the information with defined responsibilities, sources and periodicity?		
	What are the likely costs of such data collection and analysis (dollar costs in terms of the time of evaluation staff, programme managers and staff, and partners)?		
	What kind of information do the key stakeholders request?		
	What kind of information on women's rights is accessible and how it can be collected?		
Conduciveness of the context	Is the context conducive to conduct the evaluation, both external and internal to the programme, including the stakeholder's implication?		
	Are there resources available to undertake the evaluation such as well-trained staff, financial resources, equipment?		
	Do evaluation capacities and expertise exist to undertake the evaluation from a gender equality and human rights perspectives?		

Adapted from the UNDF for Women, Evaluation Guidance Note Series No.4

11.2 Prioritizing interventions for Impact Evaluation

Prioritizing interventions for impact evaluation should consider the relevance of the evaluation to the organisational or development strategy; its potential usefulness; the commitment from senior managers or policy makers to using its findings; and/or its potential use for advocacy or accountability requirements. It is also important to consider the timing of an impact evaluation. When conducted belatedly, the findings come too late to inform decisions. When done too early, it will provide an inaccurate picture of the impacts (i.e., impacts will be understated when they had insufficient time to develop or overstated when they decline over time) (betterevaluation.org).

11.3 Planning and Managing

Planning and managing an impact evaluation include:

- Describing what needs to be evaluated and developing the evaluation brief
- Identifying and mobilizing resources
- Deciding who will conduct the evaluation and engaging the evaluator(s)
- Deciding and managing the process for developing the evaluation methodology
- Managing development of the evaluation work plan
- Managing implementation of the work plan including development of reports
- Disseminating the report(s) and supporting use

Determining causal attribution is a requirement for calling an evaluation an impact evaluation. The design options (whether experimental, quasi-experimental, or non-experimental) all need significant investment in preparation and early data collection, and cannot be done if an impact evaluation is limited to a short exercise conducted towards the end of intervention implementation. Hence, it is particularly important that impact evaluation is addressed as part of an integrated monitoring, evaluation and research plan and system that generates and makes available a range of evidence to inform decisions. This will also ensure that data from other M&E activities such as performance monitoring and process evaluation can be used, as needed (betterevaluation.org).

11.3.1 Relevance and timeliness

Planning for an impact evaluation, and collecting data for an impact evaluation, should be initiated from the beginning of the programme. Impact evaluations should be conducted if their findings will be relevant for future planning, and in time to incorporate the findings into decision making. In practice when departments want to undertake impact evaluations and this has not been planned in advance the data may not be available.

If a programme manager has limited evaluation resources and needs to choose between implementation evaluation and impact or economic evaluation, there may be reasons for choosing implementation evaluation. For example, unless one knows that the programme is being

implemented according to design, there may be little reason to expect it to produce the desired outcomes. Results identified without understanding how they were achieved is of very little management use to a programme manager. In some cases, there are obvious reasons preventing impact and it is not worth the investment in an impact evaluation. However, a note of caution, you may get a well implemented programme that has no positive impact, and may indeed do harm.

For this reason, in many cases under the National Evaluation Plan, where an impact evaluation has been requested, in practice it has proved more appropriate to do an implementation evaluation first, and then plan thoroughly to do an impact evaluation at some point in the future.

11.3.2 Legitimacy

The legitimacy of an impact evaluation can be improved by ensuring that it considers the perspectives of different stakeholders in terms of what would be considered as successful implementation. This might include involving key stakeholders in the development of evaluation questions and the evaluation design, or involving the programme management team in interpreting observation and interview data. This can include beneficiaries, e.g. involving them in the process of sharing their experiences of service delivery through interviews or surveys, or involving them in the process of collecting data, through community score cards, or participatory mapping processes, or the methods of Appreciative Inquiry and Most Significant Change outlined earlier.

Legitimacy comes from explicit and transparent criteria of data extraction and analysis against explicit criteria of internal and external validity, and adequacy of reporting.

11.3.3 Credibility of the evidence

As in all evaluations, impact evaluations should be explicit about the methods chosen, the reasons for their use, their limitations and how these have been addressed. Key issues to address in terms of credibility are: the quality of existing data; the quality of additional data collected; and sampling. There is also an issue of design bias with some people believing that only RCTs, or some other evaluation design, are able to provide credible evidence of impact. In practice it is often difficult to undertake RCTs for many complex policy issues, and other methodologies are needed, which must still be carried out with rigour. More generally, all evaluation designs carry a risk of bias. Consequently, all evaluation reports should include a risk of bias assessment, and an indication of the degree to which this risk was, or was not, overcome.

Impact evaluations need to assess the quality of existing data used, such as programme reports, media reports, existing photographs and performance indicators. The methods for collecting primary data need to be carefully chosen and implemented appropriately. In particular the expertise and independence of those collecting data needs to be assessed. It is important to check whether data have been collected, and sometimes verified, by an independent agency. Additional data collection should be supported by a combination of expert knowledge about the programme and well-planned and carefully documented data collection, interpretation and analysis.

Data sources for impact evaluations should be chosen so that they triangulate important issues and balance out the limitations of any one source. Sampling decisions should be transparent, and the sampling of informants, sites and time periods should be carefully done to ensure adequate coverage, and any limitations carefully noted.

11.3.4 Trade-offs

There can be critical trade-offs for different types of impact evaluation designs. A longer intensive design that collects data from all sites may provide answers to every single evaluation question yet it may have high costs. A short, internal evaluation may be cost effective and provide answers to all the posed evaluation questions, yet lack credibility because it did not have an external evaluator.

11.4 Managing Impact Evaluation

11.4.1 Who undertakes the evaluation

Impact evaluations need to be undertaken by an independent evaluator/evaluation team that specialises in research and evaluation to ensure that the right decision is made in conducting an impact evaluation, that the appropriate methods and approaches are chosen, and that the evaluation is conducted in a technically proficient manner.

11.4.2 Terms of reference

In line with DPME Guideline for the development of ToR, the ToR for impact evaluation requires a clear understanding of the intervention as a prerequisite for the design. Sector and area expertise may not be essential but are certainly an advantage.

If the impact evaluation is defined in terms of causal attribution, the ToR should stress the need for credible counterfactual analysis. The concept notes/proposal should make clear how this will be addressed, being explicit about the evaluation approach. The evaluation team should have technical competences to implement these methods.

11.4.3 Data Sources

Good quality data are essential for impact evaluation. The evaluation design must be clear on sources of data and should also be realistic about how long it will take to collect and analyse primary data. There are various sources of data that impact evaluations can use depending on what is being evaluated and the type of approach chosen to undertake the respective evaluation. For example, data about program activities, outputs and outcomes. Other data required by the impact evaluation can depend on the methodology used. Data on other factors that may affect the outcome of interest may be needed to control for outside influences. This aspect is particularly important when using evaluation methods that rely on more assumptions than randomized methods do. Sometimes it is also necessary to have data on outcomes and other factors over time to calculate trends, as is the case with the difference-in-differences method. Accounting for other factors and past trends also helps increase statistical power. Even with randomized assignment, data on other characteristics can make it possible to estimate treatment effects more precisely. They can be used to include additional controls or analyse the heterogeneity of the program's effects along relevant characteristics.

Use of existing data

Some existing data are almost always needed at the outset of an impact evaluation to estimate benchmark values of indicators or to conduct power calculations. Beyond the planning stages, the availability of existing data can substantially diminish the cost of conducting an impact evaluation. While existing data, and in particular administrative data, are probably underused in impact evaluation in general, the feasibility of using existing data for impact evaluation needs to be carefully assessed. Data collection is often the largest cost when implementing an impact evaluation. However, to determine whether existing data can be used in a given impact evaluation, a range of questions must be considered:

- **Sampling.** Are existing data available for both the treatment and comparison groups? Are existing samples drawn from a sampling frame that coincides with the population of interest? Were units drawn from the sampling frame based on a probabilistic sampling procedure?
- **Sample size.** Are existing data sets large enough to detect changes in the outcome indicators with sufficient power? The answer to this question depends on the choice of the outcome indicators, as well as on the results of the power calculations.

- **Availability of baseline data.** Are the existing data available for both the treatment and comparison groups prior to the rollout of the program or innovation to be evaluated? The availability of baseline data is important to document balance in pre-program characteristics between treatment and comparison groups when randomized methods are used, and critical for the implementation of quasi-experimental designs.
- **Frequency.** Are the existing data collected frequently enough? Are they available for all units in the sample over time, including for the times when the outcome indicators need to be measured according to the results chain and the logic of the intervention?
- **Scope.** Do existing data contain all the indicators needed to answer the policy questions of interest, including the main outcome indicators and the intermediate outcomes of interest?
- **Linkages to programme monitoring information.** Can existing data be linked to monitoring data on program implementation, including to observe which units are in the treatment and comparison groups, and whether all units assigned to the treatment group received the same benefits?
- **Unique identifiers.** Do unique identifiers exist to link across data sources? (worldbank.org).

11.4.4 Time and Cost

The required time for a counterfactual impact evaluation depends on whether primary data collection is involved. If it is, 18 months is a reasonable estimate time from inception to final report. If there is no primary data collection then 12 months might be feasible. The survey cost is the largest cost component of an impact evaluation. However, each context will be unique and require specific budgeting discussion and decision.

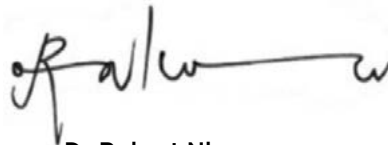
Budgeting for an evaluation is dependent on numerous factors. A general 'rule of thumb' is that an evaluation should be between 0.1% to 5% of an intervention's budget. However, this depends on many variables such as the amount of credible data already collected, the timeline to collect data, the amount of field work that needs to be done, and other contributing cost factors. Impact Evaluation has a lot of fixed and variable costs. Cost drivers include geographic scope, length of questionnaire, number of respondents, etc and fixed costs (whether done internally or hired out) include preparation of concept notes, ToRs, questionnaires, oversight, etc. while variable costs relate to the scope of the data collection.

The programme manager has a key role in ensuring that the scope of what is promised by evaluators, or expected by the programme manager, is realistic for the amount budgeted; as over ambitious and under budgeted scope of work is likely to yield a weak base of evidence and an unused report.

12. Peer review

An independent peer review should be undertaken by an independent person who is qualified in undertaking impact evaluations.

Signed

A handwritten signature in black ink, appearing to read 'R Nkuna', written over a faint, light-colored rectangular stamp or watermark.

Dr Robert Nkuna
Director-General

Department of Planning Monitoring and Evaluation
Date: 26/03/2024

Annex 2: Glossary

Impact

The positive or negative changes that result from a particular intervention or programme.

Outcome

The short, medium, or long-term results of an intervention that contribute to its overall impact.

Baseline

The initial data collection point against which subsequent changes are measured to assess impact of an intervention.

Impact Evaluation

Type of evaluation that seeks to measure changes in outcomes (and the well-being of the target population) that are attributable to a specific intervention. Its purpose is to inform high-level officials on the extent to which an intervention should be continued or not, and if there are any potential modifications needed.

Intervention

The action or process of intervening. For example, a high degree of state intervention in the economy through a programme, policy or plan.

Attribution

A concept in social psychology addressing the processes by which individuals explain the causes of behaviour and events. The problem of attribution is the problem of assigning observed changes in output and outcomes to the intervention. This is done by constructing a counterfactual.

Theory of Change

A detailed description of how and how a particular intervention is expected to lead to specific outcomes and impacts.

Comparison Group

A group of units (e.g. persons, classrooms) that receive either no treatment or an alternative treatment. The purpose of a comparison group is to serve as a source of counterfactual causal inference.

Counterfactual

Measures what would have happened to beneficiaries in the absence of the intervention, and impact is estimated by comparing counterfactual outcomes to those observed under the intervention. Outputs and outcomes in the absence of the intervention. The counterfactual is necessary for comparing actual outputs and outcomes to what would have been in the absence of the intervention.

Randomised Control Trials (RCT)

Specific type of scientific experiments, and the gold standard for a clinical trial. RCTs are often used to test the efficacy or effectiveness of various types of medical interventions within a patient population.

Quasi-Experimental design

A study design that approximate the rigor of an RCT but does not involve random assignment of participants.

Data Collection methods

Techniques used to gather data for the evaluation, such as surveys, interviews, focus groups and document analysis

Data Analysis

The process of examining and interpreting the data collected during the evaluation to draw conclusions about impact of the intervention.

Propensity Score Matching (PSM)

A statistical technique that enables evaluators to construct a counterfactual group to estimate the impact of an intervention.

Interrupted Time Series Analysis (ITSA)

It is a quasi-experimental method to establish the causal effect of an intervention.

Instrumental Variable (IV) regression

A method of estimating impact that makes use of a different variable (the instrument) to predict treatment in an econometric analysis.

Synthetic control

A quantitative method which uses historical data to construct a 'synthetic clone' of a group receiving a particular intervention.

Difference in Difference

A statistical technique used to estimate the causal effect of a treatment, policy or intervention on a particular outcome.

Qualitative Comparative Analysis (QCA)

A method used to compare different aspects of an intervention and contextual factors to understand the different characteristics or combinations of characteristics which are associated with outcomes.

Process Tracing

It is a structured method to developing and assessing theories about how a particular outcome arose.

Contribution analysis

A method which is used to understand the likelihood of whether the intervention has contributed to an outcome observed, or not commonly known as contribution claim.

Contribution Tracing

It is a rigorous mixed qual-quant participatory method to establish the validity of contribution claims in evaluation, with explicit criteria to guide evaluators in data collection and measuring confidence in findings.

Decision Tree

A tool that can help users of this guideline to decide whether or not an impact evaluation featuring causal attribution and a counterfactual is needed, or whether a different evaluation design would best suit the evaluation purpose and questions.

Selection Bias

It occurs when sample or data used in a study is not representative of the population to represent, leading to systematic differences between the characteristics of the sample and the population.

Evaluability Assessment

A systematic process that helps to identify whether a programme is in a condition to be evaluated, and whether an evaluation is justified, feasible and likely to provide useful information.

Result Based Management

A management strategy focusing on performance and achievement of outputs, outcomes and impacts.

Trade-offs

A situation that involves losing one quality or aspect of something in return for gaining another quality.

Bibliography

Babbie, E. (2013). *The practice of social research* (13th ed.). United Kingdom: Wadsworth Cengage Learning.

Babbie, E., & Mouton, J. (2001). *The practice of social research*. Cape Town: Oxford University Press.

Bamberger, M, Rugh, J, Mawbry, L, (2012): "RealWorld Evaluation: Working Under Budget, Time, Data, and Political Constraints", SAGE, 2012, 236

Bricket M et al. (2020). CECAN (2021) Magenta Book Annex A. Central Government guidance on evaluation.

CECAN (2016) Testing Contribution Claims with Bayesian Updating. A CECAN Evaluation and Policy Practice Note for policy analysts and evaluators Note No. 2.1. Winter 2016

Center for Global Development. (2006). *When will we ever learn? Improving lives through impact evaluation*. Center for Global Development.

Creswell, J. W., & Plano Clark, V. L. (2011). *Designing and Conducting Mixed Methods Research* (2nd ed.). Thousand Oaks: SAGE.

Department of Planning, Monitoring and Evaluation (2019). *National Evaluation Policy Framework*. Available: National Evaluation Policy Framework Nov 2019.pdf (dpme.gov.za)

Gertler, P. J., Martinez, S., Premand, P., Rawlings, L. B., & Vermeersch, C. M. J. (2016). *Impact Evaluation in Practice*. <http://www.worldbank.org/pdt>

Glewwe, P., & Todd, P. (2022). *Impact Evaluation in International Development: Theory, Methods, and Practice*. Washington, DC: World Bank. Doi:10.1596/978-1-4648-1497-6. License: Creative Commons Attribution CC BY 3.0 IGO

Henning, E., Van Rensburg, W., & Smit, B. (2004). Finding your way in qualitative research. Pretoria: Van Schaik.

HM Treasury (2020) Central Government Guidance on evaluation: HM Treasury – available: www.gov.uk/official-document

Krueger, R. A., & Casey, M. A. (2000). Focus groups: A practical guide for applied research. Thousand Oaks: Sage.

Neuman, W. L. (2006). Social Research Methods. Qualitative and Quantitative Approaches. Toronto: Pearson.

Nutley, S., & Powell, A. (2013). What counts as good evidence? Provocation paper for the alliance for useful evidence. www.alliance4usefulevidence.org

OECD. (2002). Glossary of Key Terms in Evaluation and Results Based Management. p37.

Patton, M. Q. (2002). Qualitative Research & Evaluation Methods (3rd Ed.). Thousand Oaks: SAGE.

Patton, M.Q. (2008): “Utilization Focused Evaluation”, Fourth Edition, Thousand Oaks, SAGE.

Rogers, P. (2012) “Introduction to Impact Evaluation”. Interaction. <http://www.interaction.org/impact>

Rossi, P.H. et al (2013). Evaluation: A Systematic Approach (7th Ed). SAGE Publication

Guideline on Impact Evaluation

Created March 2014

Updated: March 2024