Evaluation design

An evaluation design describes how data will be collected and analysed to answer the Key Evaluation Questions.

There are different pathways for you as manager depending on who will develop the evaluation design. In most cases your evaluator will develop the evaluation design. In some cases you will – if you have evaluation expertise and/or the evaluation design has already been developed (for example, in an evaluation that is intended to match an earlier evaluation).

Take into account the following important factors when developing an evaluation design

1. **The nature of what is being evaluated**
   In particular whether there are complicated or complex aspects that need to be addressed in the evaluation, and other particular challenges such as delays before impacts are evident or barriers to collecting accurate data.
   Manager's guide to evaluation :
   Consider important elements of what is being evaluated

2. **The nature of the evaluation**
   In particular the types of Key Evaluation Questions that are being asked, when the answers are needed.
   Manager's guide to evaluation :
   Consider important aspects of the evaluation

3. **Available resources and constraints**
   Resources including money, existing data, expertise, technical equipment. Constraints including requirements to use certain common indicators, limits to availability of key informants or barriers to accessing existing data.
   Manager's guide to evaluation :
   Consider the implications of the resources available and specific constraints

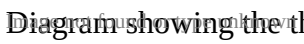

Diagram showing the three important factors flowing into appropriate evaluation methods and designs

If an EVALUATOR will develop the evaluation design

- Engage a competent evaluation expert - internal, external or a combination. (See 'Select an evaluator / evaluation team' for advice).
- Work with the expert(s) to ensure they understand important factors that should be taken into account in the evaluation design (see section above)
- The design should provide details of how data will be collected analysed. It is often useful to do this in the form of an Evaluation Matrix which shows how each Key Evaluation Question will be answered.

If YOU (as manager) will develop the evaluation design

- Understand important factors that should be taken into account in the evaluation design (see section above)
- Develop an evaluation design that addresses these important factors.
- Summarise the design in the form of an Evaluation Matrix which shows how each Key Evaluation Question will be answered.

Subsequently, arrange for a **technical review of the evaluation design** and arrange for a **review of the design by the evaluation management structure** (e.g., steering committee). Ideally this will include representation from primary intended users.

Arranging technical review of the evaluation design

Before finalizing the design, it can be helpful to have a technical review of it by one or more independent evaluators. It might be necessary to involve more than one reviewer in order to provide expert advice on the specific methods proposed, including specific indicators and measures to be used. Ensure that the reviewer is experienced in using a range of methods and designs, and well briefed on the context, to ensure they can provide situation specific advice.

Arranging review of the design by the evaluation management structure

In addition to being considered technically sound by experts, it is essential for the evaluation design to be seen as credible by those who are expected to use it.

Get formal organisational review and endorsement of the design by an evaluation steering committee (see ' Identify who will be involved in decisions and what their roles will be' for possible structures, processes and terms of reference for a steering committee)

Where possible do data rehearsal of possible findings with primary intended. This is a powerful strategy for checking the appropriateness of the design by presenting mock-ups of tables, graphs and quotes that the design might produce. It is best to produce at least 2 different versions – one that would show the program working well and one that would show it not working. Ideally the primary intended users of the evaluation will review these and either confirm the suitability of the design or request amendments to make the potential findings more relevant and credible. (For more information see Patton, MQ (2011) Essentials of Utilization-Focused Evaluation. pp. 309-321).

Consider important elements of what is being evaluated

What is being evaluated makes a difference to how it should be evaluated. It is helpful to identify particular aspects of what is being evaluated and check that these have been addressed in the evaluation design.

1. Check the stage of development of the project or program

Firstly, check the implications of the stage of development of the project or program that is being evaluated. Is it still being planned? Is it part–way through implementation? Or is it near the end – or has it in fact already ended?

Stage of development	Consequence	Possible implication for the evaluation design
Not yet started	Can set up data collection from the beginning of implementation	*Possible to gather baseline data as a point of comparison and also to establish comparison groups or control groups from the beginning*
Opportunity to build some data collection into administrative systems to reduce costs and increase coverage		
Period of data collection will be long	*Need to develop robust data collection systems including quality control and storage*	
Part way through implementation	Cannot get baseline data unless this has already been set up	*Will need to construct retrospective baseline data to estimate changes that have occurred*
Might be able to identify “bright spots” where there seems to be more success and those with less success	*Scope to do purposeful sampling and learn from particular successes and also cases which have failed to make much progress*	
Almost completed	Cannot get baseline data unless this has already been set up	*Will need to construct retrospective baseline data to estimate changes that have occurred*
Depending on timeframes, some outcomes and impacts might already be evident	*Opportunity to gather evidence of outcomes and impacts*	
Completed	Cannot get baseline data unless this has already been set up	*Will need to construct retrospective baseline data to estimate changes that have occurred*
Depending on timeframes, some outcomes and impacts might already be evident	*Opportunity to gather evidence of outcomes and impacts*	

Cannot directly observe implementation	*Will need to depend on existing data or retrospective recollections about implementation.*

2. Is it complex or complicated?

Secondly, consider whether there are important aspects that are either complicated (with many components) or complex (emergent) that should be addressed in the evaluation design.

(i) Focus

Does everyone share the same objectives?

Homogeneity of objectives	***Implications***
Everyone shares a single set of objectives	*Impacts to be included can be readily identified from the beginning.*
There are different objectives valued by different stakeholders.	*Need to identify and gather evidence about multiple possible changes*
(competing objectives, different objectives at different levels)	*Need an agreed way to weight or synthesise results across different domains to produce a judgement of overall performance.*
The stated objectives are changing (often in response to changing needs or opportunities)	*Need nimble impact evaluation systems that can gather adequate evidence of emergent intermediate outcomes or impacts*

(ii) Management

Who has responsibility for management and decision making?

Who is responsible	***Implications***
Single organisation	*Primary intended users and uses easy to identify and address in the development of Key Evaluation Questions and endorsement of the design*

Multiple organisations (which can be identified) with specific, formalized responsibilities	*Likely to need to negotiate access to data and ways to link and co-ordinate data* *Might need to negotiate parameters of a joint impact evaluation, including negotiating scope and focus.*
Changing list of organizations working together in flexible ways	*Need nimble impact evaluation systems that can gather evidence about the contributions of emergent actors and respond to the different ways they value intended and unintended impacts*

(iii) Consistency

How much variability is there in how activities are implemented?

Level of variability	***Implications***
Standardized – one-size-fits-all program	*Quality of implementation should be investigated in terms of compliance with 'best practice'.*
Adapted – variations of a programme planned in advance and matched to pre-identified contextual factors.	*Quality of implementation should be investigated in terms of compliance with the practices prescribed for that type of situation.*
Adaptive – evolving and personalised program that responds to specific and changing needs.	*Quality of implementation should be investigated in terms of how responsive and adaptive service delivery was.*

(iv) Necessity

How many different options are there for solving the problem or producing the intended impacts? To what extent is this exact initiative needed to solve the problem?

Number of possible interventions	***Implications***
There is only one way to achieve the intended impacts.	*Counterfactual reasoning appropriate.*

The intervention is one of several ways of achieving the impacts, and the options can be identified.	*Counterfactual reasoning not appropriate as it does not accept a causal relationship between the intervention and the impacts unless they would not have occurred in the absence of the intervention.*
Possibly one of several ways of achieving the intended impacts (uncertain).	*Counterfactual reasoning not appropriate as it does not accept a causal relationship between the intervention and the impacts unless they would not have occurred in the absence of the intervention.*

(v) Sufficiency

To what extent will the problem be solved by the intervention alone?

Generalisability of the intervention	***Implications***
The intervention is enough to produce the intended impacts. Works the same for everyone.	*Counterfactual reasoning appropriate* *Reasonable to ask 'Does it work?'*
Works only in specific contexts which can be identified (eg implementation environments, participant characteristics, support from other interventions).	*Impact evaluation question needs to be 'For whom, in what circumstances and how does it work?'* *Counterfactual reasoning only appropriate if the causal package of supportive context and other activities can be identified and included.*
Works only in specific contexts which are not understood and/or not stable.	*Impact evaluation question needs to be 'For whom, in what circumstances and how does it work?'* *Counterfactual reasoning not appropriate as the causal package of supportive context and other activities is changing and/or poorly understood and cannot be adequately identified.*

Change trajectory

How are the impact variables expected to change over time? For example, straight line of increase, or J curve? To what extent are the relationships between variables understandable and predictable?

Relationship between variables	***Implications***

Simple relationship (cause and effect). Predictable.	*Measurement of change can be done at a convenient time and confidently extrapolated*
Complicated relationship that needs expertise to understand and predict.	*Timing of the measurement of changes should be undertaken when it will be most meaningful – expert advice will be needed.*
Emergent factors and multiple causes, sudden changes (tipping points) that are unpredictable. Can only be understood in retrospect.	*Changes will need to be measured at multiple times as the change trajectory cannot be predicted.*

Unintended impacts

To what extent are unintended impacts predictable?

Predictability of unintended impacts	***Implications***
Easily predictable and therefore can be readily included in the data collection plans	*Need to draw on previous research and common sense to identify potential unintended impacts and gather data about them*
Need expertise to predict and address.	*Need advice from experts about potential unintended impacts and how these might be identified.*
Unpredictable - only identified and addressed when they occur.	*Need to include a wide net of data collection that will catch evidence of unexpected and unanticipated unintended impacts.*

Source: Resource Hub for Evaluating C4D 2016 - adapted from Funnell and Rogers (2011), pp.90-91, Rogers 2016.

3. Identify issues to be addressed

Are any of the following issues present? They will need to be addressed in the design.

Issue	**Possible implications for the evaluation design**

Long time until impacts will be evident	Might need to gather data about intermediate outcomes (that will be evident during the timeframe of the evaluation) and use other research and evaluation evidence to predict the likely achievement of impacts
Difficulty observing implementation activities (eg conflict affected or remote areas)	Might need to gather data through remote sensing, key informants, big data or crowdsourcing
Difficulty observing results (outcomes, impacts) (eg sensitive issues, private behaviour)	Might need to gather data through key informant interviews, or unobtrusive measures (for example looking at patterns of wear from foot traffic) or techniques for gathering sensitive data (for example polling booth)

Consider important aspects of the evaluation

Evaluations are designed to answer the Key Evaluation Questions. Different types of questions need different methods and designs to answer them.

In evaluations there are four main types of questions:

Descriptive questions ask about what has happened or how things are – for example:

- What were the resources used by the program directly and indirectly?
- What activities occurred?
- What changes were observed in conditions or in the participants?

Causal questions ask about what has contributed to changes that have been observed – for example:

- What produced the outcomes and impacts?
- What was the contribution of the program to producing the changes that were observed?
- What other factors or programs contributed to the observed changes?

Evaluative questions ask about whether an intervention can be considered a success, an improvement or the best option and require a combination of explicit values as well as evidence – for example:

- In what ways and for whom was the program successful?
- Did the program provide Value for Money, taking into account all the costs incurred (not only the direct funding) and any negative outcomes.

Action questions ask about what should be done to respond to evaluation findings – for example:

- What changes should be made to address problems that have been identified?
- What should be retained or added to reinforce existing strengths?
- Should the program be refunded?

Key Evaluation Questions often contain more than one type of questions – for example to answer the KEQ "How effective has the program been?" requires answering:

Descriptive questions – What changes have occurred?

Causal questions – What contribution did the intervention make to these changes?

Evaluative questions – How valuable were the changes in terms of the stated goals – taking into account the types of changes, the level of change and the distribution of changes.

Check the adequacy of the design by disaggregating each KEQ into the different types of questions and then checking them against the following points.

(i) Checking the adequacy of the design for descriptive questions

The design should make it clear how descriptive questions will be answered. These descriptive questions might relate to:

- Inputs – materials, staff
- Processes – implementation, research projects
- Outputs – eg research publications
- Outcomes – eg changes in policy on the basis of research
- Impacts – eg improvements in agricultural production

It can be helpful to set this out in a table that shows how data will be collected and analysed to answer these descriptive questions.

Descriptive question	Existing data that can be used	Additional data collection/retrieval	Sampling Analysis
What has been the level of resources used for the program?			
Who has participated in the program?			
What changes have occurred in terms of [specific behaviour]?			

The narrative should explain the choices made, addressing:

- Making maximum use of existing data – including a review of the quality and relevance of this
- Appropriate sampling – whether of people, sites, organisations or time periods – what type of sampling has been chosen and why this is appropriate for the type of generalization that will be undertaken.
- Appropriate data collection methods – why these methods have been chosen
- Appropriate data analysis methods – why these methods have been chosen

(ii) Checking the adequacy of the design in terms of evaluative questions

Many evaluations do not make explicit how evaluative questions will be answered – what the criteria will be (the domains of performance), what the standard will be (the level of performance that will be considered adequate or good), how different criteria will be weighted. A review of the design could check each of these in turn:

- Are there clear criteria for this evaluative question?
- Are there clear standards for judging the quality of performance on each criterion?
- Is there clarity about how to synthesize evidence across criteria? For example, is it better to have some improvement for everyone or big improvements for a few?
- Are the criteria , standards and approach to synthesis appropriate? What has been their source? Is further review of these needed? Who should be involved?

Ideally an evaluation design will be explicit about these, including the source of these criteria and standards. They might be set out in a table such as the following.

Table 1: Example table setting out the evaluative criteria, standards, synthesis process and sources

Evaluative aspect	**Process for developing agreed standards, criteria and synthesis**	**Criteria**	**Standards**	**Synthesis/Weighting**
Adequacy of resources for the program	Using national standards for the provision of services	Number of [services] per 100,000 people	[x] per 100,000 people	Average across all regions, weighted for population
Quality of services provided	National Service Standards	Financial accessibility	All people able to access services regardless of ability to pay	
Cleanliness	Food handling surfaces free from contamination			
Community consultation	Cultural appropriateness	People from all ethnic backgrounds feel welcome in the service		

(iii) Checking the adequacy of the design in terms of causal questions

Many evaluations do not make clear how causal questions will be answered. There are many designs and methods that can be used, but they involve one or more of these strategies:

(a) Compare results to an estimate of what would have happened if the program had not occurred (this is known as a counterfactual).

This might involve creating a control group (where people or sites are randomly assigned to either participate or not) or a comparison group (where those who participate are compared to others who are matched in various ways). Techniques include:

- *randomised controlled trials (RCTS)* – a control group is compared to one or more treatment groups

- *matched comparisons* - participants are each matched with a non-participant on variables that are thought to be relevant. It can be difficult to adequately match on all relevant criteria

- *propensity score matching* – creates a comparison group based on an analysis of the factors that influenced people's propensity to participate in the program

- *regression discontinuity* - compares the outcomes of individuals just below the cut-off point with those just above the cut-off point.

(b) Check for consistency of the evidence with the theory of how the intervention would contribute to the observed results

This can involve checking that intermediate outcomes have been achieved, using process tracing to check each causal link in the theory of change, identifying and following up anomalies that don't fit the pattern, and asking participants to describe how the changes came about.. Techniques include:

- *contribution analysis* – sets out the theory of change that is understood to produce the observed outcomes and impacts and then searches iteratively for evidence that will either support or challenge it.

- *key informant attribution* – asks participants and other informed people about what they believe caused the impacts and gathers information about the details of the causal processes

- *qualitative comparative analysis* - compares different cases to identify the different combinations of factors that produce certain outcomes

- *process tracing* - a case-based approach to causal inference which focuses on the use of clues within a case (causal-process observations, CPOs) to adjudicate between alternative possible explanations. It involves checking each step in the causal chain to see if the evidence supports, fails to support or rules out the theory that the program or project produced the observed impacts

- *qualitative impact assessment protocol* – combines key informant attribution, process tracing and contribution analysis, using interviews undertaken in a way to reduce biased narratives

(c) Identify and rule out alternative explanations

This can involve a process to identify possible alternative explanations (perhaps involving interviews with program sceptics and critics and drawing on previous research and evaluation, as well as interviews with participants) and then searching for evidence that can rule them out.

While technical expertise is needed to choose the appropriate option for answering causal questions, as manager you should be able to check there is an explicit approach being used, and seek technical review of its appropriateness.

Causal relationship (between one variable and another – one step in the causal chain)	What strategies and methods/designs are being used for causal inference
eg Participation in program and improved health and wellbeing	Counterfactual – matched comparison groups of participants and non-participants
eg Increased skills and changed behavior	Consistency of evidence and ruling out alternatives – process tracing and key informant attribution

(iv) Check that the design and process answers the action components of KEQs

Answers to action questions are often made in the form of recommendations. These don't necessarily flow straight from the findings. They often need an additional step of identifying possible actions and selecting the most appropriate, given the particular values and the availability of resources.

As manager you should check there is an explicit process for developing and reviewing recommendations, with appropriate levels of input from key stakeholders.

Consider the implications of the resources available and specific constraints

Identify the resources that can be used for the evaluation, and any particular constraints for them.

The following potential resources could be used for the evaluation:

- Funding to engage external individuals or organizations to design and/or conduct the evaluation or review the design and the final report
- Staff time to either conduct the evaluation or to manage an external contractor
- Time and goodwill of other stakeholders who will be involved in the evaluation – such as partner organizations, community members.
- Existing data

Identify any particular constraints for the evaluation such as:

- Short time before findings are needed to inform decisions
- Poor reputation of evaluation due to previous experiences
- Difficulties in engaging particular groups or in working collaboratively
- Missing baseline data
- Difficulties in observing or getting data about implementation or results – for example, when it is being implemented in remote locations, or in fragile, conflict-affected areas.
- Disagreement about what success looks like – for example:
    - Disagreement about the overall goals – for example, is an early childhood program primarily about improving workforce participation of parents or about early learning of children?
    - Disagreement about the criteria that should be used – for example, is good research technically very accurate or produced in time to inform an important decision? Is the goal to improve the average health and wellbeing in a community or to ensure everyone is above the minimum

requirement?
    - Diagreement about the standards that should be used - for example, is a 10% increase in published research a good result?

Do an estimate of the costs to collect and analyse the data, as well as the project management and reporting time needed. If available resources are not adequate for the design, adjust the design and/or resources.

When reducing costs it is essential to consider the implications and how to manage these risks. Some possible options for reducing costs are shown below, along with some possible implications and ideas for managing the risks.

- Reduce the number of Key Evaluation Questions
    - Possible implications: Evaluation might no longer meet the needs of the primary intended users
    - How to manage these risks: Carefully prioritise the KEQs. Review whether the evaluation is still worth doing
- Reduce sample sizes
    - Possible implications: Reduced accuracy of estimates
    - How to manage these risks: Check these will still be sufficiently credible and useful through data rehearsal using interval estimates
- Make more use of existing data
    - Possible implications: Might mean that insufficiently accurate or relevant data are used. The cost savings might be minimal if they are not readily accessible.
    - How to manage these risks: This is only appropriate when the relevance, quality and accessibility of the existing data is adequate – need to check this is the case before committing to use them
- Embed data collection in program implementation
    - Possible implications: Might lead to a reduction in data quality
    - How to manage these risks: Ensure staff are trained and motivated to collect data properly and have sufficient time and equipment to do so
- Use fewer waves of data collection, including possibly retrospectively created baselines
    - Possible implications: Will increase the risk of inaccurate data
    - How to manage these risks: Check that retrospective baselines will be sufficiently accurate and that less frequent information on progress will be sufficient to inform decisions