HM Treasury

May/2026

Magenta Book

Central Government guidance on evaluation

Contents

Foreword

The Magenta Book is the government's central guidance on evaluation. It is a critical tool in supporting the effective development and assessment of government policies, programmes and projects.

Understanding the effectiveness of public spending and the impact of our policy interventions is critical to effective decision-making. Proportionate and high quality evaluation is a strategic necessity that ensures accountability, informs future spending decisions, and drives the iterative improvement of public services.

The strength of this updated version of the Magenta Book guidance lies in its integrated perspective. In addition to providing refreshed technical guidance for analytical specialists, new guidance on value for money evaluation, test and learn approaches to policy development, and the alignment of benefits management with monitoring and evaluation ensures the new Magenta Book can serve as a shared roadmap for evaluation across the policy, delivery, finance, and analytical professions.

By embedding a culture of high-quality evaluation across government, we strengthen our collective ability to learn and improve. The principles outlined in this publication will empower our professions to deliver better evidence, better decisions, and ultimately better outcomes for the public.

Jenny Dibden
Head of Government Social Research

Conrad Smewing
Head of the Finance Function

Susan Acland-Hood
Head of the Policy Profession

Acknowledgements

The development of the updates and new focus areas of the Magenta Book was made possible through the contributions and support of various individuals from across government and external organisations. In particular, we would like to thank the following authors and editors:

Levin Wheller
Daniela Travaglia
Nicolette Lares
Jennifer Bottomley
Alison Darlow
James Collis
Leanne Dew
Gloria Crabolu
Jo Voisey
Moria Sloan
Jess Hunt
John Galilee
Kirstine Szifris
Jack Blumenau
Raphaela Berding-Barwick
Barbora Aldlerova
Alison Higgins
Oliver Hauser
Alex Norman-Rhodes
Lauren Probert
Alec Roberts
Natalia Chivite-Matthews
Steven Finch
Lucie Moore
Jon Cooper
Andrew Luty
Rebekah Eden
Mike Daly
Marzieh Talebi

We also gratefully acknowledge the contributions of key stakeholders across government, including the Departmental Directors of Analysis, Government Social Research Heads of Profession, and the Cross-Government Evaluation Group. We also recognise the foundational work provided by the authors of previous versions of the Magenta Book.

Finally, thank you to everyone involved in the external stakeholder engagement call. We appreciate the feedback provided by stakeholders within and outside of government, as well as the external peer reviews conducted by Professors Peter John and David Parsons.

Executive summary

What is evaluation?

Evaluation is a systematic assessment of the design, implementation and outcomes of an intervention.[1, 2] It involves understanding how an intervention is being or has been implemented and what effects it has, for whom and why. It identifies what can be improved and estimates its overall impacts and cost-effectiveness.

When is evaluation useful?

Evaluation can inform thinking before, during and after an intervention's implementation. Different questions are answered at each stage:

- **BEFORE** – What can we learn from previous evaluations of similar interventions?[3] How is the intervention expected to work? How is it expected to be delivered? Are its assumptions valid? Can it be piloted and tested before full roll-out? Can roll-out be designed to maximise potential learning?

  Provides evidence that informs the intervention design, how best to implement the design and what the likely outcomes might be. Helps identify and reduce uncertainty.

- **DURING** – Is the intervention working as intended? Is it being delivered as intended? What are the emerging impacts? Why? How can it be improved? Are there unintended consequences?

  Provides evidence on the implementation of the intervention and any emerging outcomes so that it can be continually improved.

- **AFTER** – Did the intervention work? By how much? At what cost? What have we learned about its design and its implementation? Are the changes sustained?

  Provides evidence on the design, implementation and outcomes, drawing out lessons for the future and providing an assessment of the overall impact of the intervention.

1 Where an "intervention" is any policy, programme or other government activity intended to elicit a change.
2 HM Treasury (2026) *The Green Book: Central Government Guidance on Appraisal and Evaluation*. London: Crown Copyright. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/685903/The_Green_Book.pdf (Accessed: April 2026).
3 Gov.UK (2013) *What Works Network Official Website*. Available at: https://www.gov.uk/guidance/what-works-network#the-what-works-network (Accessed: April 2026).

What are its purposes?

There are two main purposes for carrying out an evaluation: learning and accountability.

Learning:

- To help manage risk and uncertainty (of the intervention and its implementation).
- To improve current interventions by providing the evidence to make better decisions (and feed into performance management and benefits realisation work).
- To gain a general understanding of what works, for whom and when, and generate examples for future policymaking.
- To develop evidence to inform future interventions.

Accountability

Government departments should be accountable and transparent to the accounting officer and other stakeholders. Evidence should be generated that can demonstrate an intervention's impact or wider outcomes. Evidence of effectiveness is also needed for spending reviews and in response to scrutiny and challenge from public accountability bodies.

What does evaluation mean in practice?

Monitoring and evaluation are closely related, and a typical evaluation will rely heavily on monitoring data. To be done well, both monitoring and evaluation should be done during the policy development stage with skilled expertise to ensure real-time evidence is available during implementation to aid decision-making. A comprehensive evaluation will typically consist of:

- Analysis of:
  - Whether an intervention is being implemented as intended.
  - Whether the design is working.
  - What is working more or less well and why.

  Together, these types of questions are typically referred to as a **process evaluation**.
- An objective test of what changes have occurred, the scale of those changes and an assessment of the extent to which they can be attributed to the intervention.

  This is typically referred to as an **impact evaluation** and is investigated through theory-based, experimental and/or quasi-experimental approaches.
- A comparison of the benefits and costs of the intervention; typically referred to as a **value for money evaluation**.

In order to fully understand an intervention's design, impact and results, all elements need to be explored.

Structure of the Magenta Book

This book looks at the types of evaluation (process, impact and value for money) and the main evaluation approaches (theory-based and experimental), as well as setting out the main stages of developing and executing an evaluation. The chapters are:

- Chapter 1: Why, how and when to evaluate?
- Chapter 2: Evaluation scoping.
- Chapter 3: Evaluation methods.
- Chapter 4: Data collection, data access and data linking.
- Chapter 5: Managing an evaluation.
- Chapter 6: The use and dissemination of evaluation findings.
- Chapter 7: Evaluation capabilities.
- Annex A: Analytical methods for use within an evaluation design.

Supplementary guides provide further detail on particular topics:

- Quality in Qualitative Evaluation.
- Realist Evaluation.
- Handling Complexity in Policy Evaluation.
- Government Analytical Evaluation Capabilities Framework.
- Guidance for Conducting Regulatory Post Implementation Reviews.
- Test and Learn Annex.
- Transparency Annex.

This Book is to be used in conjunction with the Green Book,[4] other government standards[5] and codes of conduct.

4 HM Treasury (2026) *The Green Book: Central Government Guidance on Appraisal and Evaluation*. London: Crown Copyright. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/685903/The_Green_Book.pdf (Accessed: April 2026).
5 Government Analysis Function (n.d.) *Government Functional Standard GovS 010: Analysis*. Crown copyright. Available at: https://www.gov.uk/government/publications/government-analysis-functional-standard--2 (Accessed: April 2026).

Key terminology

The table below sets out some key concepts and their terminology used in this Book (note that other, non-government guides may use slightly different wording).

Term	Use in the Magenta Book
Evaluation design	The overarching design of the whole evaluation, which includes how the evaluation will meet the learning aims specified during the scoping stage. Chapters 2, 3, 4, 5 and 6 cover elements that form the overarching evaluation design.
Evaluation types	The type of evaluation is defined by the evaluation questions (see Table 2.2). Common types of evaluation include process, impact and value for money.
Evaluation approach	The way that the answering of evaluation questions is approached; for example, impact evaluations may use a theory-based approach and/or an experimental approach.
Evaluation methods	The way that information is collected and analysed in order to test theories and answer the evaluation questions (e.g. difference-in-difference, modelling, randomised control trials)
Data collection	The collection of information to use in evaluation; this can be quantitative or qualitative.
Intervention	Anything intended to elicit change, including a programme, policy, project, regulation and changes in delivery method.
Artificial Intelligence (AI)	A branch of computer science that seeks to develop algorithms that can perform tasks traditionally requiring intelligence when performed by a human, such as learning, problem-solving and decision-making

HM Treasury

Chapter One

Why, how and when to evaluate?

1

Summary

Evaluations of government interventions should be proportionate and fit for purpose.

Evaluation plays a role in policy design, development and delivery, as well as in informing the design of subsequent interventions.

Planning an evaluation early allows for an intervention to be designed in a way that can maximise the learning that can be gained. It can also reduce the costs of data collection by building this into the intervention's delivery.

There are three main types of evaluation: process, impact and value for money evaluations; each focused on answering different types of questions. For a full understanding of whether an intervention worked, how, why, for whom and at what cost, all three types of evaluation are required.

A good evaluation is useful, credible, robust, proportionate and tailored around the needs of various stakeholders, such as decision-makers, users, implementers and the public. By responding to potential users' needs, the outputs should be both usable and useful.

Planning an evaluation requires consideration of both the design and the project management of the evaluation. This typically requires expertise and resource.

1.1 Introduction

It is essential that public money is well spent, that government intervention is well-targeted, and that any regulation is an appropriate balance between burden and protection. The government, public and all other stakeholders should be able to learn from and build on what has gone before. They should also be able to scrutinise whether the intervention was effective, the outcomes were achieved and the money was well spent. Evaluation is one way to achieve this accountability and learning. All policies, programmes and projects should be subject to proportionate evaluation.

The Magenta Book has been written for government decision-makers and government analysts to help them understand the role of evaluation and the processes and methods for conducting an evaluation. It should also benefit the wider research community, particularly those bidding for government work, and other commissioners, such as local authorities and charities, who develop and deliver policies and interventions.

Government aims to conduct proportionate, fit-for-purpose evaluations that are genuinely useful to decision-makers. In the immediate term they can provide evidence that can improve the intervention being examined. In the longer term, they can help build the evidence base, inform future policy development and delivery, and assess value for money.

Evaluation tools and methods can also be part of a broader Test and Learn way of working designed to maximise the chance of achieving intended outcome/s. See the Magenta Book Supplementary Guidance on Test and Learn for further information.[6]

6 *The Magenta Book should be read alongside the* Green Book: guidance on appraisal and evaluation in central government*, which sets out why and how to conduct appraisal of government policy, and the rationale for the early planning of evaluations.*

1.2 What is policy evaluation?

Policy evaluation is the systematic assessment of a government policy's design, implementation and outcomes. It involves understanding how a government intervention is being or has been implemented and what effects it has had, for whom and why. It also comprises identifying what can be improved and how, as well as estimating overall impacts and cost-effectiveness.

Evaluations differ in scale and ambition, but at their core they all seek evidence to answer questions, such as:

- Is the intervention working as intended?
- Is it working differently for different groups?
- Why, or why not, might it be working differently for different groups?
- How is the policy operating in practice?
- Where can the policy be improved?
- What was the overall impact of the policy?
- Is it value for money?
- If we were to do it again, what would we do differently?

1.3 Why evaluate?

Two primary reasons to evaluate are learning and accountability.

1.3.1. Learning

In terms of learning, evaluations can provide the evidence with which to manage risk and uncertainty. Especially in areas that are innovative or breaking new ground, there is a need for evidence to illustrate whether an intervention is working as intended. Early learning can also illuminate which parts are particularly successful or unsuccessful and what needs to be adapted to improve performance. Pilots can be useful in this context as they allow the design, implementation and outcomes to be tested in a controlled environment at a smaller scale to generate evidence to inform a broader policy initiative.

Even areas with less uncertainty can often benefit from evaluation: to provide evidence to inform a benefits management strategy to help realise the anticipated benefits,[7] or understand how to maximise the efficiency and effectiveness of delivery. Even when we are very confident that an intervention will be effective, we would at the very least want to monitor outcomes and confirm that they are in line with expectations.

Evaluations also generate learning on what works for whom, when and why. Interventions are rarely conducted in isolation and are typically one strand of a greater programme, building on what has gone before, and soon replaced with another idea. It is important that we learn from interventions, so that we can apply that learning to subsequent policies in the same area or other related areas. Even policies that are terminated because they are considered ineffective or too costly can produce valuable learning about mistakes to avoid in the future or identify whether any elements of the intervention were successful.

Fundamentally, learning is about good decision-making. Evaluation can provide evidence to inform decisions on whether to continue an intervention, how to improve it, how to minimise risk, or whether to stop and invest elsewhere.

7 Infrastructure and Projects Authority (2025) *The Teal Book: Guidance on Project Delivery*. London: Crown Copyright. Available at: https://projectdelivery.gov.uk/teal-book/home/ (Accessed: April 2026).

1.3.2. Accountability

Another main reason to evaluate is for accountability purposes. Government makes decisions on people's behalf and spends tax collected from individuals and businesses. Government also uses regulatory initiatives, which run the risk of being overly burdensome or having perverse outcomes for some. Government has a responsibility to maximise public value and outcomes delivered for taxpayers' money and government activity. Evaluation has a crucial role to play in this.[8]

Government departments must also inform the public about the outcomes and value of the initiatives they put in place and be accountable and transparent to their accounting officer for their spending. Evidence of intervention effectiveness is also required for spending reviews and in response to scrutiny and challenge from bodies, such as:

- National Audit Office[9]/Public Accounts Committee.
- Select Committees.
- Infrastructure and Projects Authority (IPA).
- Better Regulation Executive/Regulatory Policy Committee.
- International Development Committee.

In some cases, such as the following, evaluation is mandatory:

- Regulatory policies subject to post-implementation review (PIR).[10,11]
- Regulations containing a sunset or a duty-to-review clause.
- To meet the requirements of the International Development Assistance Act 2015 Evaluations for accountability tend to focus on monitoring and assessing impact.

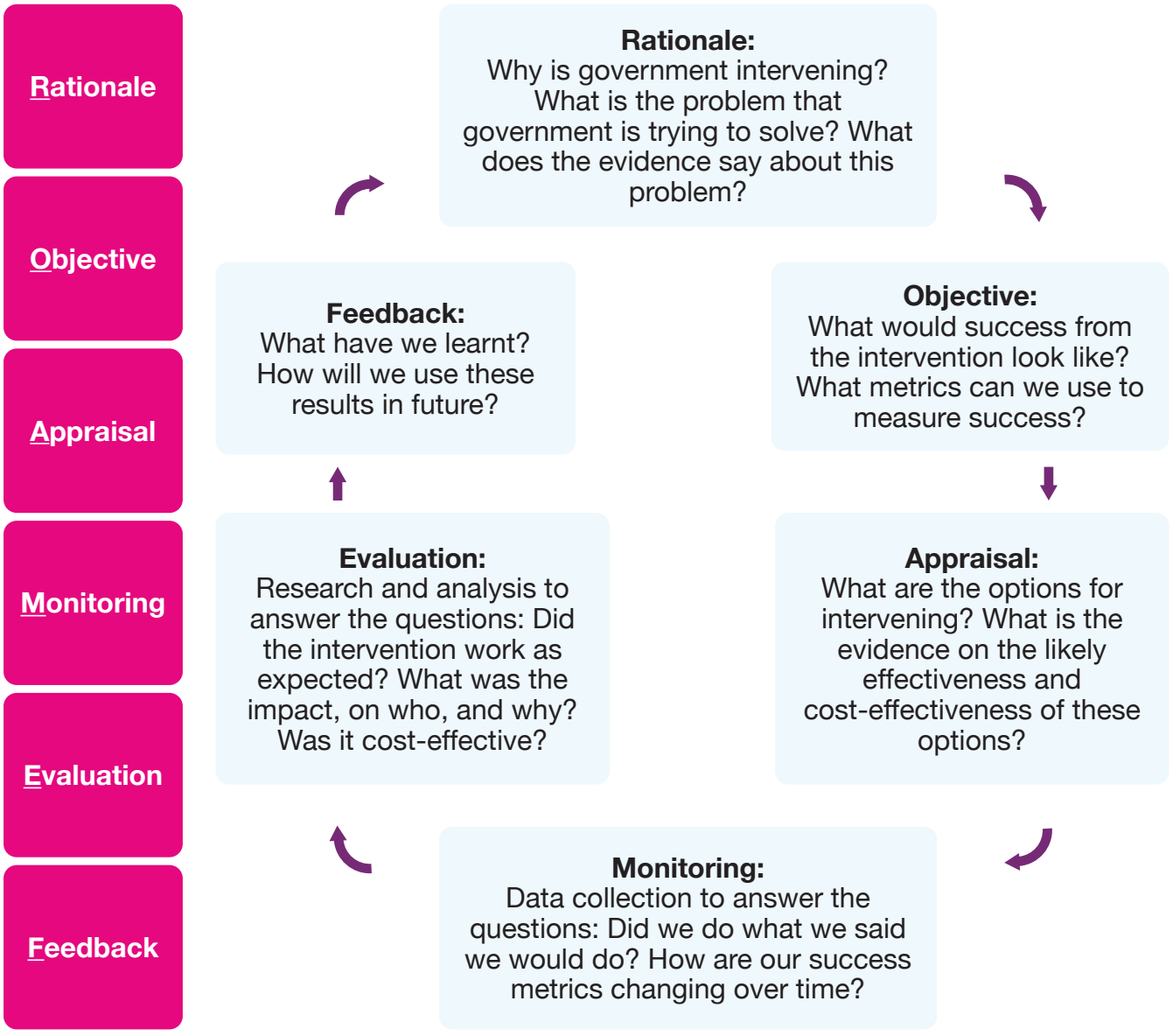
In practice, balancing learning and accountability can be difficult. There will be more or less of a focus on each depending on the role that evaluation will play with the specific intervention under consideration and the needs of stakeholders.

8 HM Treasury (2019) *The Public Value Framework*. London: Crown Copyright. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/785553/public_value_framework_and_supplementary_guidance_web.pdf (Accessed: April 2026).
9 National Audit Office (n.d.) *About us – Our work: Value for Money*. Available at: https://www.nao.org.uk/about-us/ (Accessed: April 2026).
10 Department for Business, Energy and Industrial Strategy (2023) *Better Regulation Framework Guidance*. Crown Copyright. Available at: https://assets.publishing.service.gov.uk/media/67587ba55a2e4d4b993bfa83/better-regulation-framework-guidance-2023.pdf (Accessed: April 2026).
11 Department for Business, Energy and Industrial Strategy (2023) *Small Business, Enterprise and Employment Act 2015: Statutory Guidance under s.31 of the Small Business, Enterprise and Employment Act*. Crown Copyright. Available at: https://www.gov.uk/government/publications/small-business-enterprise-and-employment-act-statutory-review-requirements (Accessed: April 2026).

1.4 What role does evaluation have?

Evaluation has a role at all stages in the intervention's life cycle. The Green Book presents a framework for the appraisal and evaluation of all interventions, programmes and projects known as 'ROAMEF'. The ROAMEF framework is useful for thinking about the key stages in the development of a proposal, from the articulation of the rationale for intervention and the setting of objectives, through to options appraisal and, eventually, implementation and final evaluation, including the feeding back of evaluation evidence back into the policy cycle.

Figure 1.1: The ROAMEF cycle



In practice, ROAMEF is a simple way of expressing a complex process. In reality, none of the steps is an isolated activity and each will inform and be informed by the other steps. It will rarely proceed as a linear process. Evaluation is useful at all stages.

The outputs and learning from earlier evaluations should be fed in at the rationale and objectives stage, when the issue to be tackled is being explored. Evaluators often take active roles in the development of well-defined objectives (such as SMART[12]), which set out exactly what changes the intervention aims to bring about and how these changes will be measured.

Commencing Theory of Change thinking (see Chapter 2) is useful at this stage to help articulate objectives and stress-test potential intervention ideas.

At appraisal stage,[13] when options to address the issue are being examined in detail, previous evaluation evidence will be invaluable to assess the feasibility and cost of these options. Early evaluation thinking and piloting can be crucial in testing policy ideas (exploring questions such as: will this work? why? how? for whom?). Theory of Change work can help articulate how various options are expected to work and the strength of the evidence that underpins them. It will become clear what data are available and where uncertainties and risks lie. It is at this stage that evaluation planning should start in earnest, so that the intervention and the evaluation can be designed in parallel to provide the evidence required to meet the learning and accountability objectives.

Evaluation evidence is useful when designing a new intervention or reviewing an existing policy. How useful this proves to be is dependent on how tailored the design of the evaluation is to the needs of decision-makers. Iteration is common, and early learning from monitoring and evaluation can result in speedy changes to the policy design and objectives. A 'Test and Learn' approach to evaluation design is becoming more popular, with fast feedback loops taking place to influence the intervention design and delivery (see Figure 1.2).

Test and Learn activity focuses on refining specific elements of an intervention, particularly those where uncertainty is highest, by testing assumptions about feasibility, behaviour, implementation and mechanisms in real-world settings. Its aim is to iteratively improve these components – using appropriate social research methods – so that the final design of the intervention has the greatest chance of success. The primary purpose at this stage is learning and adaptation, not definitive judgement about overall impact. Crucially, this iterative activity does **not** replace the need for robust evaluation of the intervention as a whole.

Once a design has been strengthened through successive rounds of testing and decision-making, a prospective evaluation – such as a randomised controlled trial or other counterfactual method – may still be required to establish whether the optimised intervention delivers its intended outcomes and represents good value for money when implemented at scale. Further evaluation during wider roll-out may also be needed to understand how the intervention performs in different places, systems and populations.

Test and Learn is most valuable where policy problems or delivery contexts are complex or uncertain, and where key assumptions about behaviour, feasibility or outcomes

12 SMART stands for Specific, Measurable, Achievable, Realistic and Timebound. By describing objectives in this way, it makes it easier to communicate the aims of the intervention. This can help the evaluator later.
13 Department for Transport (2016) *Strengthening the Links between Appraisal and Evaluation*. London: Crown Copyright. Available at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/540733/strengthening-the-links-between-appraisal-and-evaluation.pdf (Accessed: April 2026).

are weakly evidenced. It is particularly relevant when there are several plausible ways to intervene, when outcomes are influenced by multiple factors, organisations or behaviours, or when interventions can be developed and implemented in stages. In these circumstances, Test and Learn supports small-scale testing, rapid learning and iterative refinement before significant investment or scale-up.

By contrast, it is less suitable where problems and solutions are well understood, requirements are fixed, or delivery must follow a fully specified, non-iterative approach. Used appropriately, Test and Learn helps manage uncertainty early and strengthens the basis for robust, proportionate evaluation later in the policy cycle.

Further practical guidance on applying a Test and Learn approach is provided in the full annex.

Figure 1.2: How Test and Learn uses proportionate evaluation tools across the ROAMEF cycle

EXPLORE	Build a shared understanding of how the system operates, who is affected and what shapes behaviour and outcomes, analysing data, and identifying the riskiest assumptions to test.
CO-DESIGN	Collaboratively develop and prototype potential solutions with frontline staff, users, partners, and other stakeholders who have relevant insight into the defined problem; test early assumptions and refine the Theory of Change.
TEST	Trial the most promising prototypes in real-world settings at small scale, using mixed social research methods to refine design, test mechanisms, and address feasibility before wider rollout.
GROW	Use robust evaluation types (process, impact, value for money) to inform decisions about whether to scale the intervention more widely or to adapt it to specific places and contexts where different conditions may apply.

1.5 When to evaluate?

Evaluation can often be thought of as something that happens after an intervention has been implemented. However, evaluation should inform thinking throughout the ROAMEF cycle – before, during and after implementation – and has maximum utility if thought about in this way.

Before an intervention is fully formed, evaluation should be used to help shape its design and how it will be implemented. Using existing evaluation evidence, working through the Theory of Change (Chapter 2), piloting and early testing of policy ideas can explore:

- How the intervention is expected to work and what evidence supports this thinking.
- Why the intervention might not work and what evidence is there to support this.
- Where the risks and uncertainties lie.
- How the intervention works at a smaller scale and in a controlled environment.
- What baseline evidence should be used to measure future change against.

During implementation is when there is the greatest opportunity for the evaluation to influence decisions and help ensure the intervention can realise its intended benefits. During implementation, evaluations will typically look at gaining evidence about the efficacy of the intervention's design, its implementation and emerging outcomes. It can examine questions such as:

- Is the intervention being delivered as intended?
- Is the intervention working as intended?
- Is it working in the same way in different areas and for different target groups?
- What are the early indications of possible effect size?
- How is it operating in practice?
- Are there any unintended consequences?
- Can design or delivery be improved ('in-flight adjustments')?

After an intervention has been implemented, the entire intervention can be examined looking at more conclusive statements on the design, implementation and outcomes, answering questions such as:

- Did the intervention work?
- What were the effect size and cost (and was it different for different groups)?
- What contribution did the intervention make to the outcome?
- How did this relate to what was predicted in the appraisal?
- Were there any unintended or negative impacts?
- Does this represent value for money?
- What have we learned about what works in this space? What are the transferable lessons?

1.6 Why start planning early?

Building evaluation into an intervention's design is a critical way to ensure that:

- The evaluation delivers useful findings to those designing and implementing policies.
- The evaluation is of appropriate quality for its intended use.
- The right data are collected in the most cost-effective and efficient way possible.

The relationship between an intervention's design and evaluation usefulness is crucial. Small changes in an intervention's design can make the difference between a high-quality, useful evaluation and one that is not able to answer the key questions under consideration (does it work, to what extent, for whom and why?). These design changes can be large – for example, by allocating the intervention randomly to establish a treatment and control/comparison group as would happen with a randomised control trial – or small, for example ensuring that individual-level administrative data is available for analysis purposes, by establishing mandatory questionnaires to be completed by all recipients of a product or service, or by selecting the optimum sequencing in which a programme is rolled out.

It is particularly important to plan early what data and evidence should be collected before implementation and during the lifetime of the intervention. If these activities are left until after the lifetime of the intervention, it may limit the ability to conduct appropriate evaluation. For example, when adopting a Test and Learn approach, the EXPLORE and TEST phases could help establish and refine data and evidence needs.

Additionally, it may be necessary to find an appropriate comparison group or collect baseline data before implementation so that a counterfactual can be estimated. It is also the case that recipients may be more amenable to be interviewed while the programme is live and delivery bodies may only exist for the lifetime of the programme.

1.7 Aligning evaluation with business planning

To be in line with the Treasury approvals process for programmes and projects,[14] existing evidence and evaluation needs should be identified in the strategic business case with an initial estimate of resources required to meet evaluation requirements (see Figure 1.3 below). Within the 5-case model for business cases, monitoring and evaluation plans should be detailed in the management case. For regulatory initiatives, a similar approach should be taken, developing the monitoring and evaluation plans into the impact assessment process. To ensure coordination and efficiency, a holistic approach to monitoring, benefits management and evaluation should be taken. For regulatory initiatives, this may mean defining the scope of the evaluation to be wider than the statutory instrument in question, broadening it to look at a wider view of the policy being changed. Throughout, evaluators should work closely with other analysts, finance and policy colleagues to ensure that evaluation is embedded as a core component of the programme or project plan (for example, in the management case). Crucially, appraisal and evaluation plans should reflect each other, ensuring that the costs and benefits appraised can be measured and evaluated. They should also ensure that financial provision for the evaluation is included in the economic and financial cases.

14 HM Treasury (2018) *Guide to Developing the Project Business Case*. London: Crown Copyright. Available at: https://www.gov.uk/government/publications/business-case-guidance-for-projects-and-programmes (Accessed: April 2026).

Figure 1.3: Evaluation planning expected at each appraisal stage.

Strategic Business Case / Consultation Impact Assessment An initial assessment of the monitoring and evaluation activity required	**Include:** • A decision on the scale of a proportionate evaluation. • Likely scale of data requirements. • An initial estimate of budgets and resources required. • A discussion of the scope for piloting and testing prior to implementation.
Outline Business Case / Final Impact Assessment A more developed picture of the scale of ambition and emerging plans	**Include:** • The main objectives of monitoring and/or evaluation. • Likely uses and users. • A high level timetable setting out the main stages of evaluation planning and delivery against policy timetables. • Outline of budget and resources.
Full Business Case / Enactment Impact Assessment Firms up scope of work to ensure sufficient resources are secured and data collection responsibilities are clear	**Include:** • Detail on who will be responsible for data access arrangements with delivery partners. • An initial set of evaluation questions. • Detail of likely research required. • The budget and resources required for monitoring and evaluation.

1.7.1. Relationship between benefits management and evaluation

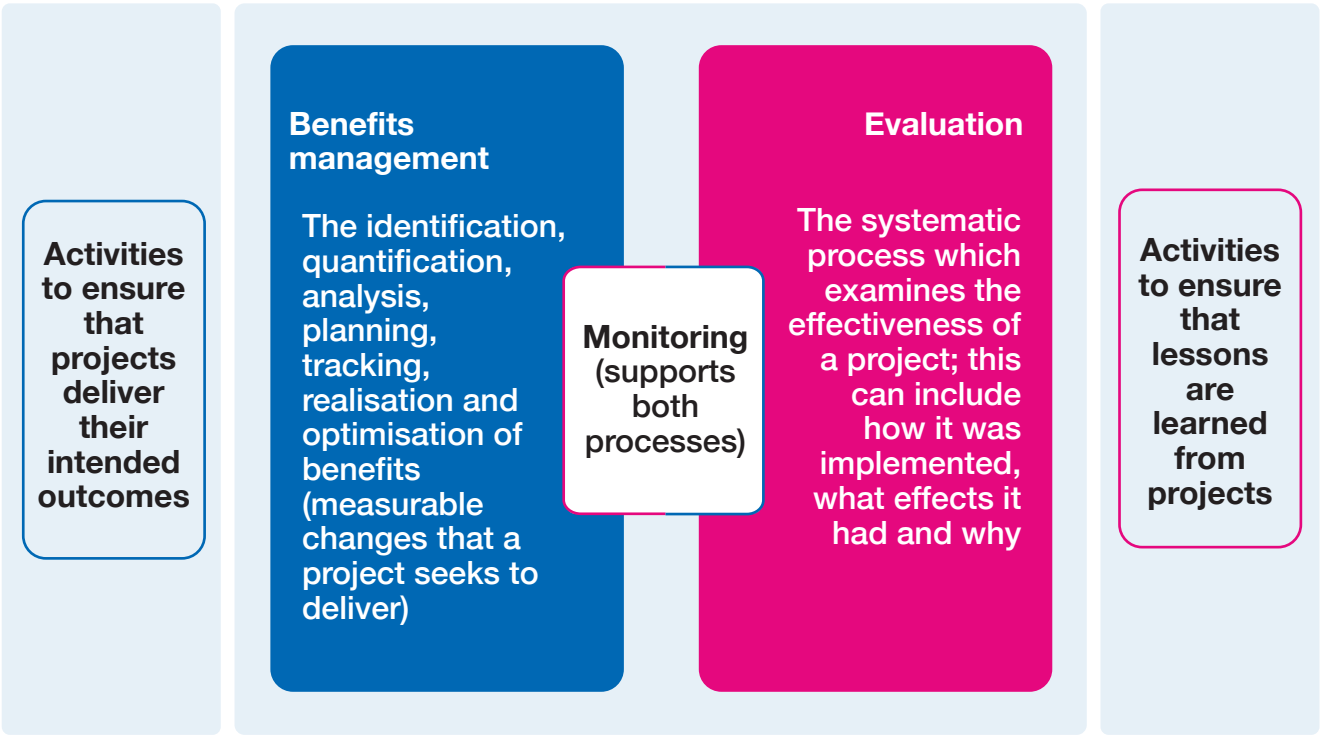
Benefits management and evaluation are complementary disciplines that can work together to define and measure the success of an intervention. Benefits management is a project delivery discipline that focuses on defining, measuring and then realising the desired changes or benefits of an intervention, while minimising any negative impacts (or 'disbenefits').[15] It is conducted and evolves throughout the project life cycle and involves continuous monitoring to track progress. Benefits management can involve planning and change activities to try and ensure the intended benefits are realised (this process can be referred to as 'benefits realisation'). It does not concern itself with why, how or where the change was derived from.

Evaluation complements the benefits management process by providing a broader assessment of whether the intervention worked as intended, often exploring the 'how' and 'why' behind the observed outcomes results and examining additional or unintended impacts beyond the anticipated benefits. An evaluation may also look at wider outcomes or impacts which cannot be directly managed, such as any longer-term economic or social impacts of the intervention.

An evaluation is often needed to understand the *additional* impact of the intervention, by considering what would have happened in the absence of the policy ('the counterfactual') and assessing the extent to which any observed changes can be attributed to the intervention.

15 Infrastructure and Projects Authority (2022) *The Teal Book: Project Delivery in Government*. Available at: https://projectdelivery.gov.uk/teal-book/home/ (Accessed: April 2026).

Figure 1.4: Diagram showing the relationship between benefits management and evaluation



Early in the project life cycle, a decision will need to be made about how benefits management and evaluation can be used to understand the outcomes of a project. In most cases it will be appropriate to use a mix of benefits management and evaluation techniques, with benefits management covering monitoring of performance measures and an evaluation study providing analysis of wider benefits, unintended consequences, attribution of change to or contribution to outcomes from the intervention and evidence on the additionality of the outcomes or benefits. Evaluation can also test whether the assumed medium and longer- term impacts will occur from the observed outputs or outcomes.

In practice, monitoring, benefits management and evaluation approaches may interact and inform one another in different ways at different stages of the policy cycle. It is therefore important that they are carefully planned at an early stage and implemented in a joined-up manner to maximise opportunities for learning.

1.8 Types of evaluation

There are three main types of evaluation activity: process evaluation, impact evaluation and value for money evaluation. These approaches are not mutually exclusive but are complementary, often crossing over to provide a comprehensive understanding of an intervention.

1.8.1. Process evaluation

'What can be learned from how the intervention was delivered?'

Process evaluations tend to examine activities involved in an intervention's implementation and the pathways by which the policy was delivered. These might vary quite considerably according to the nature of the intervention and will be policy-specific, covering questions such as:

- What worked well and less well, and why?
- What could be improved?
- How has the context influenced delivery?

Process evaluations typically use a wide range of methods, both quantitative and qualitative. They will often cover subjective issues (such as perceptions of how well a policy has operated) and objective issues (the factual details of how an intervention has operated, typically using administrative data, where available).

1.8.2. Impact evaluation

'What difference has an intervention made?'

Impact evaluations focus on the changes caused by an intervention; measurable achievements which either are themselves, or contribute to, the objectives of the intervention.

Typical impact evaluation questions include:

- What measurable outcomes, both intended and unintended, occurred?
- How much of these outcomes can be attributed to the intervention?
- Have different groups been impacted in different ways, how and why?
- How has the context influenced outcomes?
- Can the intervention be reproduced?

1.8.3. Value for money evaluation

'Is this intervention a good use of resources?'

While impact evaluation demonstrates and quantifies outcomes, it cannot on its own assess whether those outcomes are justified. Value for money evaluation considers such issues, including whether the costs of the policy are outweighed by the benefits. It can also assess whether the intervention remains the most effective use of resources.

Value for money evaluation assesses social (or public) value. Assessing social value involves comparing any costs and benefits that affect the welfare or wellbeing of the population.

The Green Book defines 'value for money' as a balanced judgement of the best use of public resources to achieve a given set of objectives.[16] That judgement involves considering monetisable social benefits and social costs, which can be expressed in monetary terms and summarised in a benefit-cost ratio (BCR). However, value for money assessment should also consider any significant unmonetisable social benefits and social costs which cannot be expressed in monetary terms, and therefore cannot be included in a BCR.

The Green Book provides more detailed guidance on social cost-benefit analysis and the valuation of non-market impacts.[17]

A standard value for money evaluation will compare the costs and benefits achieved through the programme against the original expectations outlined at the appraisal stage. These expectations will normally be set out in a business case or impact assessment. A more developed value for money evaluation could also compare the benefits and costs of different ways of achieving a given strategic objective.

Typical value for money evaluation questions include:

- What are the benefits and costs of the intervention?
- How do the costs and benefits of the intervention compare to what was expected in the appraisal or business case?
- How cost-effective was the intervention?
- How does the ratio of costs to benefits compare to that of alternative interventions, or to doing nothing?
- How economic, efficient, effective and equitable was the intervention?[18]
- Is the intervention the best use of resources?

16 HM Treasury (2022) *Green Book Supplementary Guidance: Value for Money*. Available at: https://assets.publishing.service.gov.uk/media/62443d2c8fa8f5277b365ad7/Green_Book_supplementary_guidance_-_Value_for_Money.pdf (Accessed: April 2026).
17 HM Treasury (2026) *The Green Book: Appraisal and Evaluation in Central Government*. Available at: https://www.gov.uk/government/publications/the-green-book-appraisal-and-evaluation-in-central-government#fnref:44 (Accessed: April 2026).
18 For definitions of each term, please see HM Treasury (n.d.) *The Magenta Book: Guidance on Evaluation in Central Government*, Annex A.

1.9 What is a 'good' evaluation?

There are no set criteria for defining a good evaluation; it will be determined by many factors relating to the intervention, the use the evaluation evidence will be put to, and the design and execution of the evaluation itself. These factors should be considered from the outset of evaluation planning and be made clear in the dissemination of evaluation findings.

In terms of the nature of the intervention and the use the evaluation evidence will be put to, important factors to consider in evaluation design are:

- The nature of the system before the intervention is introduced, for example, the complexity of the existing policy landscape, population size, type of market failure, stakeholder interests and so on.
- The scale, complexity, and level of innovation of the intervention.
- Existing evidence on the intervention and, therefore, the uncertainty involved.
- The nature of the intervention: what are the likely consequences and for whom?
- The timing and nature of key intervention decisions.
- The scope to change the intervention at these times.
- The status and value of the intervention.
- New or novel interventions that are currently unevidenced.

A good evaluation is fit for purpose: it is proportionate in scale and reflects the needs of decision-makers and those scrutinising the intervention from the outside.

General principles that can guide decisions around evaluation to maintain high quality:

1. **Useful**: An evaluation is high quality when it is designed to meet the needs of the many stakeholders involved (see section 1.10) and produces useful, usable outputs at the right point in time. Tailoring evaluation around known decision points and areas of policy debate are key to achieving this goal. In addition, clear communication of the limitations of evaluation findings must always be given to ensure results are used responsibly.
2. **Credible**: To be useful, evaluations need to be credible. This is often achieved through ensuring a degree of objectivity. This can be achieved through the evaluation work being conducted by an independent group of evaluators, or through respected independent figures steering and peer reviewing both the design and outputs of the evaluation work. Transparency is crucial for credibility.

3. **Robust**: Although there are no objective criteria for quality, an evaluation should be well designed, with an appropriate evaluation approach and methods, and well executed (for example, ensure adequate sampling strategies and sample sizes in surveys to identify statistically significant change; achieve sufficient power in experimental designs; qualitative sampling that ensures a wide range of voices are heard; adequate assessment of the uncertainty of model inputs and outputs; and adherence to ethical principles). The approach to establishing impact should involve rigorous comparison either in time, between groups or to alternative theories. Independent peer review and independent steering can help quality-assure the design and execution of an evaluation.
4. **Proportionate**: Proportionality is a key concept in evaluation. Not all interventions will require the same level of scrutiny or have the same learning needs. In the case of a low-risk, well-evidenced and low-priority intervention, a light-touch monitoring and evaluation exercise to ensure it has been delivered as intended and achieved the predicted outcomes is likely to be all that is necessary. On the other hand, a high-risk, high-status policy breaking new ground is likely to require a large-scale evaluation. Criteria for 'priority' interventions that require substantial evaluations typically include:

    a) High profile policies.
    b) High levels of uncertainty/risk (including possible negative consequences).
    c) High cost (if evaluating a pilot, the full cost of rolling out the policy should be considered).
    d) High learning potential (low-priority interventions on other criteria can have a high potential for filling strategically important evidence gaps).

1.10 Who are the stakeholders of evaluation?

At its core, evaluation should be useful to its many different audiences and users. Building the evaluation design around users' needs will ensure they are engaged and that the outputs are of use to and used by them.

In order for evaluation to meet its twin goals of accountability and learning, there are a number of different 'customers' whose needs should be taken into consideration.

1. **Those responsible for the intervention under consideration**: the people who have most to gain from evidence that can reduce risk and uncertainty, and from learning what is working and what is not.
2. **Those responsible for future policies**: this group will require evidence on what worked (and/or did not), why and how, and on transferable lessons.
3. **Those responsible for appraisal analysis**: they will have the most insight into what evidence and data were missing from the appraisal of the intervention, and what will be useful for the appraisal of future policies.
4. **Those responsible for scrutinising government decisions and spend**: those that hold government to account are an eager audience for evidence around the efficacy of the intervention's design and delivery, and its impact and cost.
5. **Participants/recipients of the policy**: those affected by the policy are typically also key participants in the evaluation. Their input is required, but they will also have evidence needs and a perspective on what elements of the policy should be focused on.
6. **Those delivering the intervention**: typically, although policies are often designed in central government, they are delivered by others, in many cases through a long delivery chain. Evaluation should be alive to the needs and issues of all those in the delivery chain.
7. **The public (often via the media)**: a key line of accountability is to the public who are keen to know that government money is being spent wisely, and that we are learning from past experience.
8. **Academics/other researchers**: government is rarely the only interested party in a specific policy area. Academics and other researchers are often able to spend time scrutinising government data. It is important to work with them to ensure the best use of the research evidence is being made and the maximum learning is being extracted.

Building an evaluation's design around the users of the service, potential users of the evaluation and the intended purpose of the evaluation evidence will help maximise the usefulness and use of the evaluation findings.

1.11 Who conducts evaluation?

Evaluation, although a simple concept – in essence, learning from implementation – can be a very technical and complex task to do well. There is a large, global evaluation community, and several specialist techniques and conceptual frameworks can be used. As a result, evaluation is something that is best designed, overseen and managed by individuals and teams with specialist expertise.

In government, evaluation is typically designed by specialist evaluators or analysts from the social research, economics, statistics and operational research professions, in collaboration with those designing and implementing an intervention. Often, the actual evaluation itself – collecting existing and new data, and analysing and interpreting findings – is contracted out to independent specialists (see Chapter 5).

1.12 The stages of an evaluation

Figure 1.5 below illustrates the main stages in an evaluation. Table 1.1 provides more detail about each stage, showing the key evaluation steps and the project management steps that should accompany them. Each stage is explored in more detail in Chapters 2–6 which follow.

Figure 1.5: Overview of the evaluation process

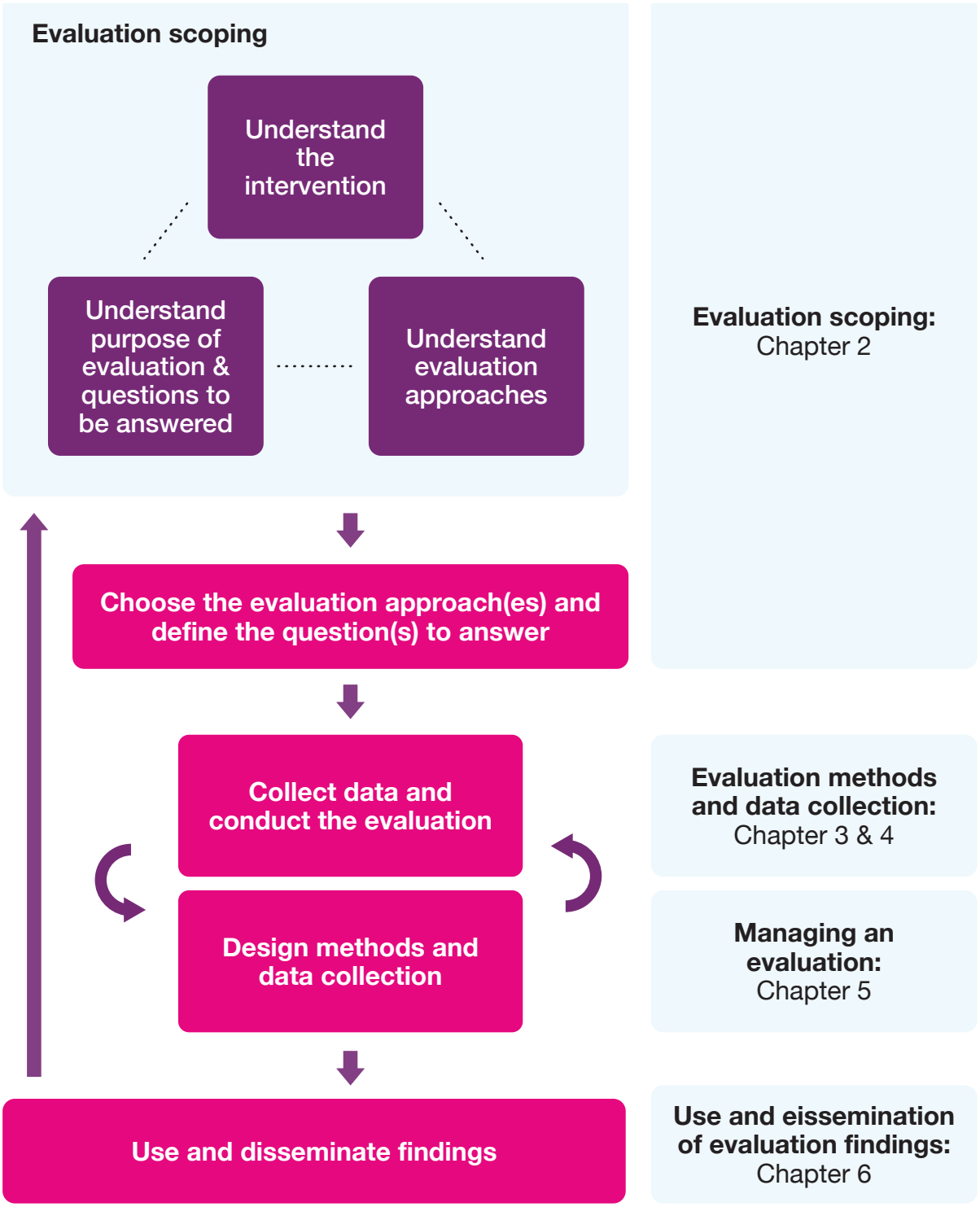


Table 1.1: Stages in planning and executing an evaluation

Stage	Chapter	Key evaluation steps	Key project management steps
Evaluation scoping	2	• Understand the intervention, what it aims to achieve, by when and for whom? (develop SMART objectives). • Understand the evidence base surrounding the intervention. • Develop the Theory of Change. • Understand the questions to be answered.	• Identify decision-points, learning goals and the type of evaluation needed. • Review programme documentation and have discussions/workshops making use of problem structuring methods where appropriate.
Evaluation design	2	• Identify the evaluation approach(es) that will help meet the learning goals. • Begin to plan the evaluation, deciding on the design and questions to be answered and the reporting points where evidence is needed.	• Agree governance, funds, timetable and method of delivery (internal/externally commissioned). • Understand whether it is possible to make changes to the policy design to improve evaluation design. • Agree required outputs and timings. • Agree evaluation questions with main stakeholders. • Quality-assure (QA) your design. • Consider approach to peer review.
Choose the appropriate methods	3	• Decide on the methods, both for analysis and data collection, that can answer the evaluation questions. • Ensure the chosen methods complement each other and are as efficient as possible.	• Estimate the cost of the chosen methods. • Estimate timings and likely reporting points.

Stage	Chapter	Key evaluation steps	Key project management steps
Conduct the evaluation	4 & 5	• Execute the evaluation, modifying design in response to learning and policy changes/stakeholder needs. • Feed in evidence where possible, in line with known and new decision points.	• Draft and quality assure (QA) specification for the work. • Commission external elements of the work. • Agree ongoing QA process. • Agree dissemination plans. • Maintain tight oversight of work. • Feed in progress and emerging outputs to policy/ programme board. Review progress and modify where necessary.
Disseminate, use & learning	6	• Prepare final evaluation analysis and outputs.	• Implement end-of-project QA plan. • Implement dissemination plan, including publication. • Work with policy/other stakeholders to utilise learning. • Conduct lessons learned exercise.

1.13 Openness and transparency

Openness and transparency are central to producing evaluation evidence that is useful, credible and robust. Transparency serves to strengthen the integrity of, and trust in, evaluation findings. Transparency also maximises value for money and promotes efficiency by enabling the reuse of code, data and other evaluation materials, and promotes effective relationships with suppliers. Ultimately, transparency in evaluation ensures that policy decisions are underpinned by evidence which is credible, trustworthy and open to scrutiny, enhancing the confidence of political audiences and the public in government evaluations.

Transparency and openness applies across the whole evaluation process depicted in (see Figure 1.5 above). This includes three core parts of the evaluation process:

- **Planning** transparency (documenting the intent, design and methodology of an evaluation before it begins).
- **Analysis** transparency (preserving any materials, data and code used to generate an evaluation's findings and, for quantitative evaluations, ensuring that final results are computationally reproducible).
- **Reporting** transparency (publishing all evaluation reports and fully describing all key methodological choices, communicating the uncertainty around findings, and demonstrating the robustness of results to alternative analytical approaches).

As a guiding principle, evaluations produced by government should be as open as possible and as closed as necessary. This means that openness should be the default, but that the level of transparency in government evaluation research should be proportionate to the value, risks, costs, and effort involved in each individual evaluation. In practice, this means that greater transparency should be applied to evaluations which:

- Are high cost.
- Evaluate high-profile or high-cost policy interventions.
- Have high learning potential or potential for re-use.
- Use novel or complex methodological approaches.

Legal constraints and constraints relating to data sensitivity may also shape the choice of transparency practices. Evaluators should therefore exercise informed judgement to determine the appropriate level of transparency for each project or programme.

Detailed guidance on the application of transparency practices can be found in the Transparency in Government Evaluation Research (TIGER) annex.

HM Treasury

Chapter Two

Evaluation scoping

2

Summary

This chapter sets out the first step of an evaluation: the scoping of an evaluation. Evaluation scoping is an iterative process that looks at:

- Defining the intervention.
- Identifying the evaluation needs.
- Understanding the most appropriate evaluation approaches.

Defining the intervention includes identifying and synthesising existing evidence and developing a Theory of Change. This should be the first step in any evaluation.

Identifying the evaluation needs involves working with the potential users of the evaluation to understand their key questions and how the evidence will be used.

The most appropriate approach to evaluating your policy or programme will depend on the type of evaluation and questions that need to be answered.

This stage of evaluation planning is iterative. Decisions made throughout evaluation planning should always be reviewed considering the purpose of the evaluation and needs of users.

2.1 Introduction

Evaluations, designed carefully alongside the design of an intervention, can produce timely, tailored evidence before, during and after implementation and delivery.

Figure 2.1, Scoping, designing and conducting the evaluation, illustrates the stages involved in an evaluation. These stages should be iterative as highlighted in Figure 1.3, Overview of the evaluation process, with a strong interaction between the learning goals, the emerging findings and more practical issues around decision points and available funds.

Although iterative, the main early stages of evaluation are:

Evaluation scoping

1. Understanding the intervention and its current evidence base.
2. Understanding the type of evaluation required. This depends on the purpose of the evaluation: to understand the **process** by which the intervention was implemented; the **impact** of the intervention; or the **value for money** of the intervention.
3. Understanding the appropriateness of the various evaluation approaches. Suitability will depend on the specific questions the evaluation will aim to answer, as well as the feasibility of the approach (see section 2.2.4 for further detail on identifying the most suitable approach to impact evaluation) Impact evaluation approaches, has further detail on identifying the most suitable approach to impact evaluation).

Evaluation design

1. Agree the most appropriate evaluation approach.
2. Identify the most appropriate method(s) to be used. Within each evaluation approach, there are a multitude of methods that could be employed. Evaluators must use the scoping work, alongside information on the availability of data, resources and timescales in making their choice (see Chapter 3).

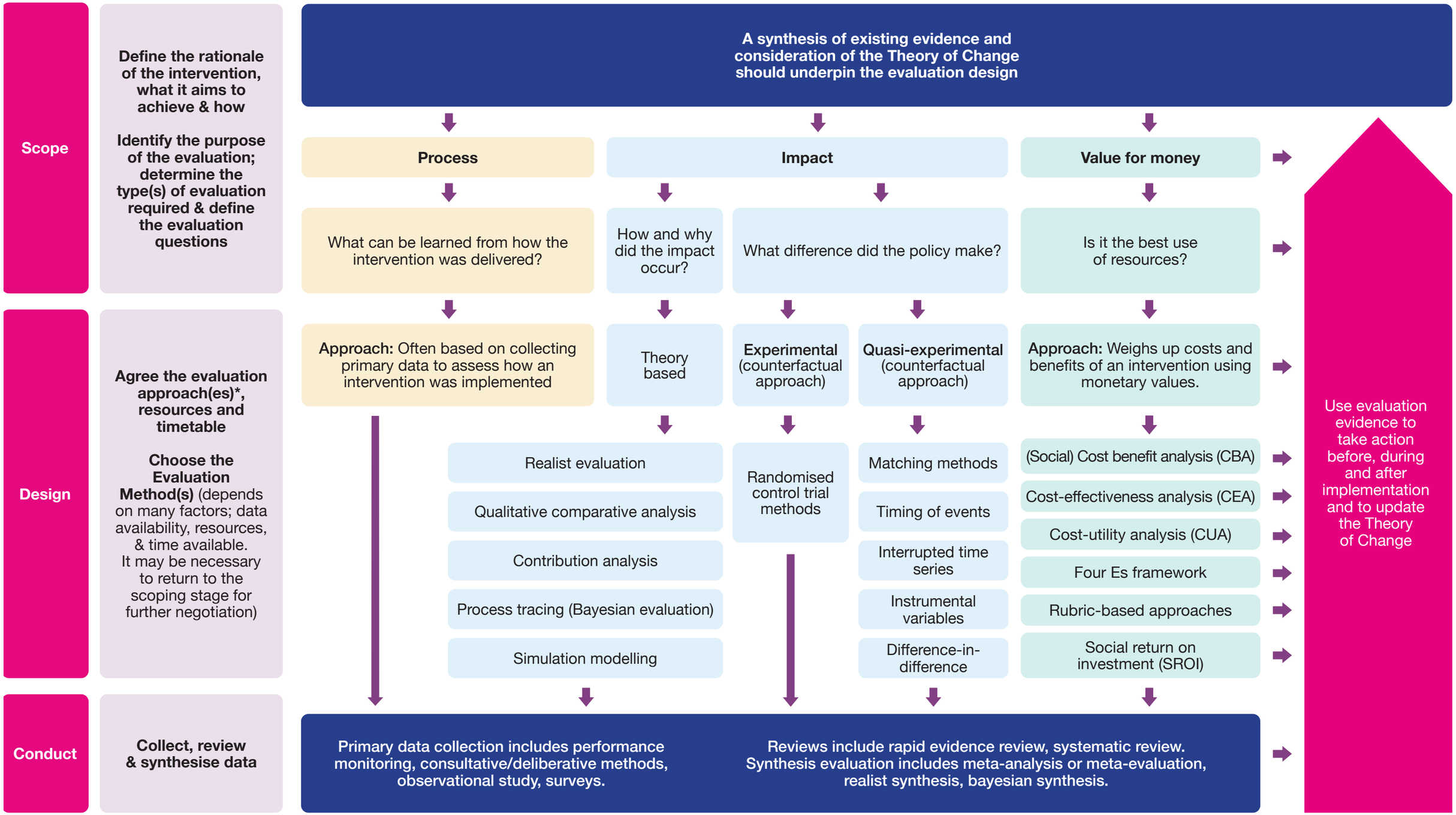
Chapter 2 focuses on the first stage, evaluation scoping, working through each step described above. Chapter 3 describes the common evaluation methods and their pros and cons. Both chapters should be worked through before reaching the final step, evaluation design.

Figure 2.1 below should be read from top to bottom. It describes the steps in stages, thus:

1. The **scoping stage**: where the evaluator determines the purpose of the evaluation and therefore, the type of evaluation required (process, impact or value for money), as well as the evaluation question(s) to be answered.
2. The **design stage**: where the evaluator chooses the most suitable evaluation approach and then the method to be used. In Figure 2.1 below, an example of some of the methods that could be used under each approach are given.
3. **Conducting the evaluation**: where various data collection, research, and review and synthesis methods may be employed to support the evaluation method chosen.

It is important to note that, although different evaluation approaches will answer different evaluation questions, they can often be complementary. For example, the evidence produced by a process evaluation can be useful input for theory-based impact evaluation methods, such as contribution analysis. Likewise, value for money evaluations generally rely on impact evaluation findings. These could include quantitative estimates of an intervention's impact produced using experimental or quasi-experimental methods.

Figure 2.1: Scoping, designing and conducting an evaluation



* Many evaluations can also be participatory or emancipatory (e.g. using developmental & action research methods). These are particularly useful in complex settings (see Complexity Guide).

2.2 Evaluation scoping

Scoping an evaluation is a crucial step. Taking the time to understand the intervention, the questions that need to be answered and the types of evaluation that can be used to answer these questions will result in an evaluation that is more likely to be of genuine use.

2.2.1. Understand the Theory of Change

Good policymaking necessitates a thorough understanding of the intervention and how it is expected to achieve the expected outcomes. Good evaluation also requires this understanding. Thoroughly examining the proposed intervention ensures:

- An understanding of how the intervention is expected to work in practice, such as the problem the intervention aims to address; the change it aims to bring about; the causal chain of events that are expected to bring about the change; the main actors; the groups expected to be impacted; and the expected conditions required for the intervention to succeed.
- Exposing the assumptions upon which the intervention is based and the strength or weakness of the evidence supporting these assumptions.
- An examination of the wider context, such as other policy changes or changes in economic, social and environmental factors.
- Designers and implementers of the intervention have the opportunity to stress-test the intervention design and ensure they agree on how the intervention is expected to work.

Understanding the intervention is typically done through synthesising existing evidence and producing a **Theory of Change**. A Theory of Change captures all of the detail listed above, including the theory of how the intervention is expected to work (setting out all the steps expected to be involved in achieving the desired outcomes), the assumptions made, the quality and strength of the evidence supporting them, and wider contextual factors.

A key part of producing a Theory of Change is the **synthesis of existing evidence**. By bringing together and assessing the strength of existing evidence about the intervention, evaluators can begin to see where evidence is weaker. This can help to identify key questions the evaluation will need to answer.

Review and synthesis of existing evidence

Synthesis of evidence can be used throughout an evaluation but is fundamental to the scoping stage. Identifying what is already known can help reduce the scale of the planned evaluation and focus it on particular areas of uncertainty.

Synthesis can be light-touch or extensive. Where evidence synthesis has been conducted during the development of the appraisal and business case for the intervention, this should be where to start.

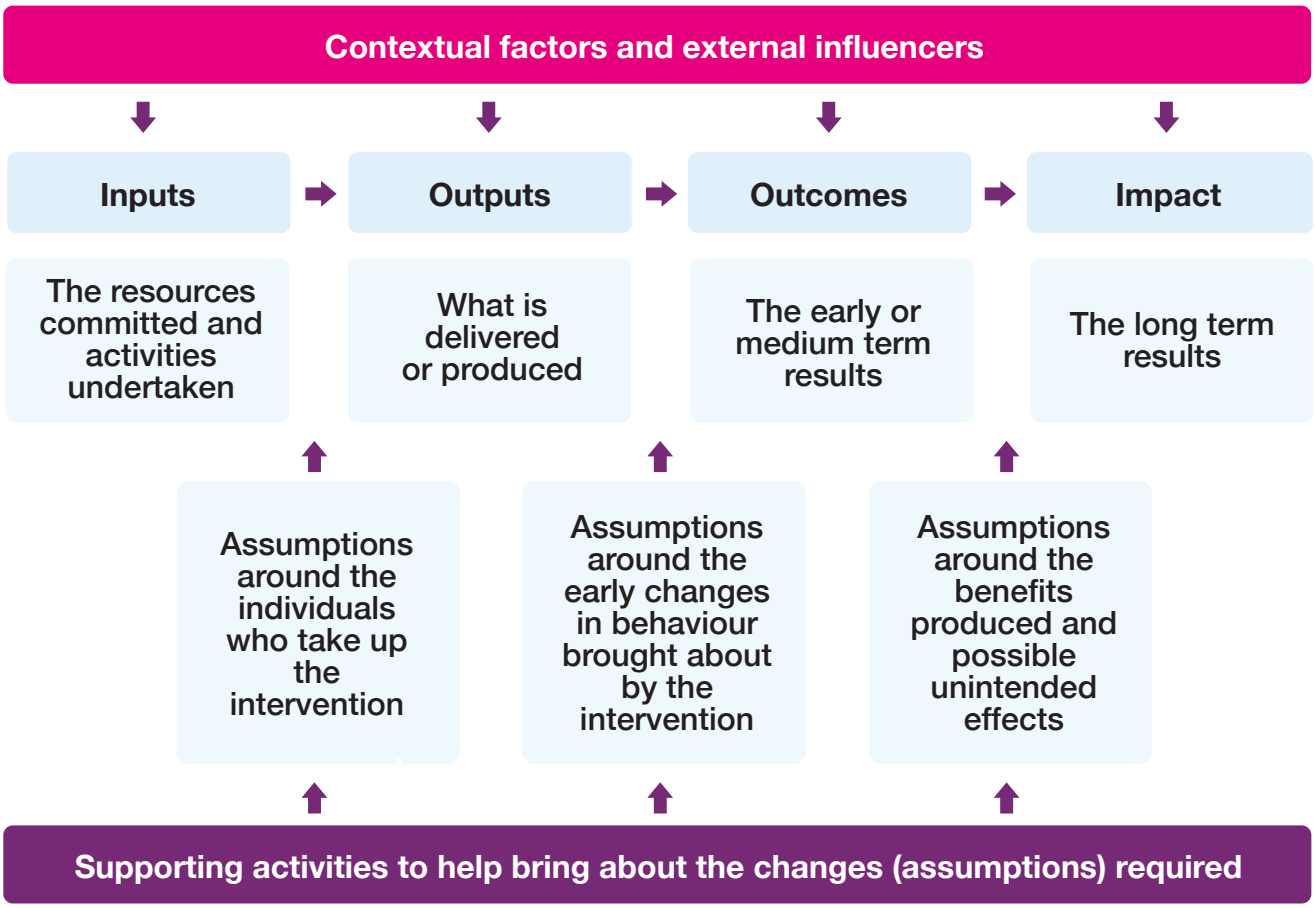
An evidence synthesis can increase the value of an individual study's findings by using meta-analysis (or meta-evaluation) to draw together findings from different studies and demonstrate replicability. Where studies appear to conflict, the evaluator should be alert to the possibility that the intervention may work for some groups in some circumstances, but not for others, and this might point to areas to explore further.

Synthesis can evolve as the evaluation proceeds, to ensure that emerging findings are put into context as the evidence base develops. For a programme of evaluation activities, an over- arching framework of outcomes or questions can prepare for the later synthesis of the various streams of evidence/evaluation activities.

Developing a Theory of Change typically involves considering the proposed inputs (what investment/regulation/actions will take place) and the causal chain that leads from these inputs through to the expected outputs and outcomes (see Figure 2.2).[19] It considers the causal mechanisms by which an intervention is expected to achieve its outcomes, basing this theory on the gathering and synthesis of evidence.

19 While logic models and diagrams (such as Figure 2.2) are essential for visualising causal chains in an intervention, it is important to recognise that these diagrams are representations of the overall Theory of Change. Government Analysis Function (n.d.) *The Theory of Change Process – Guidance for Outcome Delivery Plans.*
Available at: https://analysisfunction.civilservice.gov.uk/policy-store/the-analysis-function-theory-of-change-toolkit/ (Accessed: April 2026).

Figure 2.2: Example of a linear Theory of Change (based on Mayne 2017[20])



There are many mapping tools that can be used to help explore the intervention and how it is expected to work, often described as the 'programme theory'. These include Theory of Change mapping, logic model, outcomes mapping and system mapping.[21] The most appropriate tool to use will depend on the characteristics of the intervention, the complexity of the system it is applied to, and the type of evaluation that is being planned.

Developing a Theory of Change will typically involve the stakeholders involved in designing and executing the intervention. This can be through workshops or consultations. Alongside this, research methods, including evidence synthesis, focus groups and expert panels, can be used to gather and synthesise evidence to use in its development.

Theories of Change can range from simple descriptions to more complex analyses. More sophisticated exercises produce a more detailed and rigorous assessment of the intervention and its underlying assumptions. They detail the precise causal mechanisms that lead from one step to the next, alternative mechanisms to the same outcomes, and the assumptions behind each causal step; the evidence that supports these assumptions; and how different contextual, behavioural and organisational factors may affect how,

20 Mayne, J. (2017) 'Theory of change analysis: Building robust theories of change', *Canadian Journal of Program Evaluation / La Revue canadienne d'évaluation de programme*, 2(2), pp. 155–173. doi: 10.3138/cjpe.31122.
21 All of these processes simply involve different ways of (but are not limited to) mapping causes and effects, often in a chain formation (for example, as depicted in the 'input, output, outcome and impact' chain in Figure 2.2, Example of a linear Theory of Change).

or if, the outcomes come about. It can also be useful to explore the negative programme theory – all the reasons why the causal steps might not happen in practice and why the desired outcome is likely not to occur. This can help identify risks and issues to explore as part of the evaluation. In more complex interventions, the Theory of Change may capture interactions, feedback loops and an analysis of system boundaries.

The best way to develop the Theory of Change is through extensive collaboration with a wide range of stakeholders, including designers, implementers, beneficiaries and/or interest groups to understand how the intervention is likely to work from a range of perspectives. When this is not possible, at the very minimum it should be stress tested with key stakeholders to test whether it reflects their view of how the intervention is likely to work. If the intervention is large with many elements, a series of models can be usefully developed, focusing on different aspects of the intervention.

Crucially, the Theory of Change should continue to be developed over the lifetime of the evaluation as new evidence is developed.

Sources of further information on developing a Theory of Change

1. Better Evaluation (2015) *Theory of Change Thinking in Practice: Hivos Theory of Change Guideline*. COLOPHON. Available at: https://www.betterevaluation.org/tools-resources/theory-change-thinking-practice-stepwise-approach (Accessed: April 2026).

2. Mayne, J. (2017) 'Theory of change analysis: Building robust theories of change', *Canadian Journal of Program Evaluation / La Revue canadienne d'évaluation de programme*, 32(2), pp. 155–173. doi: 10.3138/cjpe.31122.

3. Funnell, C. and Rogers, J. (2011) *Purposeful Program Theory: Effective Use of Theories of Change and Logic Models*. San Francisco: Jossey-Bass Publishers.

4. Davies, R. (2016) *Evaluating the Impact of Flexible Development Interventions*. Methods Lab. Available at: https://cdn.odi.org/media/documents/10361.pdf (Accessed: April 2026).

5. Vogel, I. (n.d.) *ESPA Guide to Working with Theory of Change for Research Projects*. Directorate of the Ecosystem Services for Poverty Alleviation (ESPA) Programme. Available at: https://www.espa.ac.uk/files/espa/ESPA-Theory-of-Change-Manual-FINAL.pdf (Accessed: April 2026).

6. Better Evaluation (2017) 'Using logic models and theories of change better in evaluation' [blog]. Available at: https://www.betterevaluation.org/en/blog/Using-logic-models-and-theories-of-change-better-in-evaluation (Accessed: April 2026).

7. Hills, D. (2010) *Logic Mapping Hints and Tips for Better Transport Evaluations*. The Tavistock Institute. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/3817/logicmapping.pdf (Accessed: April 2026).

2.2.2. Understanding the evaluation purpose/questions

Evaluations can be designed to answer a wide range of potential questions. It is important to be clear from the outset what these questions are, how the findings from them are expected to be used, by whom, and when. This will inform the evaluation approach to be used, help focus the evaluation, and ensure the findings stand the strongest chance of having an impact on decision-making.

The questions to be answered by an evaluation will be informed by:

- The stated **purpose of the evaluation**.
- The questions **identified by the Theory of Change**, for example:
    - What are the areas of uncertainty?
    - Where are the important weaknesses in the evidence base?
    - What are intermediate outcomes that can measure progress towards the ultimate outcome?
- The questions that **stakeholders** (such as those funding, designing, implementing or impacted by an intervention. See section 1.10) want to have answered.
- The **early decision points** for the intervention (such as whether to continue its roll-out, whether to modify its design or implementation, and expected or planned review points) and what evidence will be needed to inform those decisions.
- How findings are expected to be used, considering both the short-term needs (such as benefits- realisation) and longer-term needs (such as answering 'what works?' and 'why?' questions to inform similar policies).

Table 2.1 below sets out some key evaluation uses.

Table 2.1: Key evaluation uses

Evaluation use	Description
Identification of risks	Evaluation activities, including the Theory of Change, can identify risks or dependencies and allow the intervention design and implementation to be modified in response.
Benefits management[22] and accountability	Frequent monitoring and evaluation outputs allow an assessment and explanation of progress towards realising the intended benefits. This enables corrective action to be taken where necessary.
Informing key decision points	Different decision points will have associated evidence requirements: the more explicit these are, the better able the evaluation is to meet them. The timing pressure this brings might then influence the evaluation design, methods chosen or the robustness of the evidence that will be deemed acceptable. Example decision points include: • Programme boards. • Legislative decision points. • Decisions on wider roll-out. • Regulatory policies subject to post-implementation review (PIR).[23,24] • Regulations containing a sunset or a duty-to-review clause. • To meet the requirements of the International Development Assistance Act 2015.

22 Infrastructure and Projects Authority (2016) *Guidance for Departments and Review Teams: Assurance of Benefits Realisation in Major Projects*. London: Crown Copyright.
Available at: https://www.gov.uk/government/publications/assurance-of-benefits-realisation-in-major-projects (Accessed: April 2026).
23 The Better Regulation Framework outlines the Post Implementation Review process. Department for Business, Energy and Industrial Strategy (2025) *Better Regulation Framework Guidance*. London: Crown Copyright.
Available at: https://www.gov.uk/government/publications/better-regulation-framework (Accessed: April 2026).
24 Statutory guidance on reviews includes guidance on when to include a review clause. Department for Business, Energy and Industrial Strategy (2023) *Small Business, Enterprise and Employment Act 2015: Statutory Guidance under s.31 of the Small Business, Enterprise and Employment Act*. London: Crown Copyright. Available at: https://www.gov.uk/government/publications/small-business-enterprise-and-employment-act-statutory-review-requirements (Accessed: April 2026).

Evaluation use	Description
To respond to external scrutiny	Government departments need evidence of policy effectiveness for spending reviews and in response to scrutiny and challenge from bodies such as: National Audit Office[25]/Public Accounts Committee Select Committees Infrastructure and Projects Authority (IPA) Better Regulation Executive/Regulatory Policy Committee International Development Committee
Communicating impact	The evaluation findings can be used in both internal and external communications.
Understanding context	Interventions may have different impacts depending on the context within which they are used. Evaluations can help unpick the relative impact of different contextual factors and the intervention itself. This can suggest whether the policy can be expected to work in other contexts.
Stakeholder engagement and empowerment	Including stakeholders in the design and implementation of evaluations can help build stronger relations and a common understanding of the issues the intervention seeks to address. It also allows the evaluation to respond to specific issues as they arise, as well as increases the likelihood that stakeholders will use and value the evaluation findings.
Future policy decision making	Monitoring and evaluation evidence contributes to the long-term evidence base, which is called upon when making fast-paced policy decisions. It also informs resource allocation events, such as Spending Reviews and business planning.

25 National Audit Office (n.d.) *About us – Our work: Value for Money*. Available at: https://www.nao.org.uk/about-us/ (Accessed: April 2026).

An evaluation is framed by the list of evaluation questions to be answered. In combination with the programme theory, they define the scope of an evaluation – and it is essential to the success of the evaluation to get them right.

Scoping the list of potential evaluation questions is likely to result in a long list. Few evaluations will be able to answer every question posed, typically because of the time and resources that would be required to answer them all, as well as methodological limits.

In practice, a bottom-up generation of questions should be accompanied by a top-down approach to generate a small number of high-level questions (typically around 6 or 7) under which the more detailed questions will sit. This keeps the evaluation manageable and helps the evaluators to remain focused on the key questions throughout their work.

Evaluation questions typically evolve during evaluation design and implementation, depending on feasibility, data availability, practical issues during the evaluation's execution, emerging findings and other considerations. For this reason, it is vital to maintain strong links with the users of evaluation, so that evaluation designs evolve with their needs in mind.

2.2.3. Understanding the type of evaluation required

The questions generated in the scoping exercise are likely to cover all three evaluation types: process, impact and value for money. The types of questions that relate to each are set out in Table 2.2. For a full understanding of an intervention, aspects of all the three types of evaluation are likely to be needed.

Alone, there will be weaknesses in each of these evaluations: for example, knowing the extent of the impact of an intervention will rarely explain why this impact occurred. This is especially important when the expected outcomes are not achieved. In this case, process evaluation can provide essential evidence to understand whether the issue is the result of the intervention design or the intervention delivery Process evaluation can also assess whether these intervention design or delivery issues can be overcome.

Table 2.2: Evaluation questions and types of evaluation[26]

Process evaluation questions: What can be learned from how the intervention was delivered?	**Impact evaluation questions:** What difference did the intervention make?	**Value for money evaluation questions:** Was this a good use of resources?
Was the intervention delivered as intended? • Were there enough resources? • Were there any unexpected or unintended issues in the delivery of the intervention? • To what extent has the intervention reached all the people that it was intended to? What worked well, or less well, for whom and why? What could be improved? What can be learned from the delivery methods used? • Could the intervention have been procured and delivered for less cost? • How has the context influenced delivery? • How did external factors influence the delivery and functioning of interventions? • How did external factors influence the attitudes and behaviours of target groups?	Did the intervention achieve the expected outcomes? • To what extent? Did the intervention cause the difference? • To what extent can the outcomes be attributed to the intervention? How confident can we be that the intervention caused the observed changes? • What causal factors resulted in the observed impacts? • How much can be attributed to external factors? • What would have happened anyway? How has the context influenced outcomes? • Has the intervention resulted in any unintended outcomes? • Have the outcomes been influenced by any other external factors? To what extent have different groups been impacted in different ways, how and why? Can the intervention be reproduced? What generalisable lessons have we learned about impact?	How cost-effective was the intervention? • Cost per unit (outcome, participant, and so on). • What were the costs of delivering the intervention? • Has the intervention been cost-effective (compared to alternatives and compared to doing nothing)? • What is the most cost-effective option? What was the value for money of the intervention? • What are the benefits? • What are the costs? • Do the benefits outweigh the costs? • What is the ratio of costs to benefits? Is the intervention the best use of resources? • How does the ratio of costs to benefits compare to that of alternative interventions?

26 Stern, E. (2015) *Impact Evaluation: A Guide for Commissioners and Managers*. London: Bond. Available at: https://www.bond.org.uk/wp-content/uploads/2022/09/impact_evaluation_guide_0515-2.pdf (Accessed: April 2026).

Process evaluation questions: What can be learned from how the intervention was delivered?	**Impact evaluation questions:** What difference did the intervention make?	**Value for money evaluation questions:** Was this a good use of resources?

Future learning. The different types of evaluation can together help answer questions used for future learning:

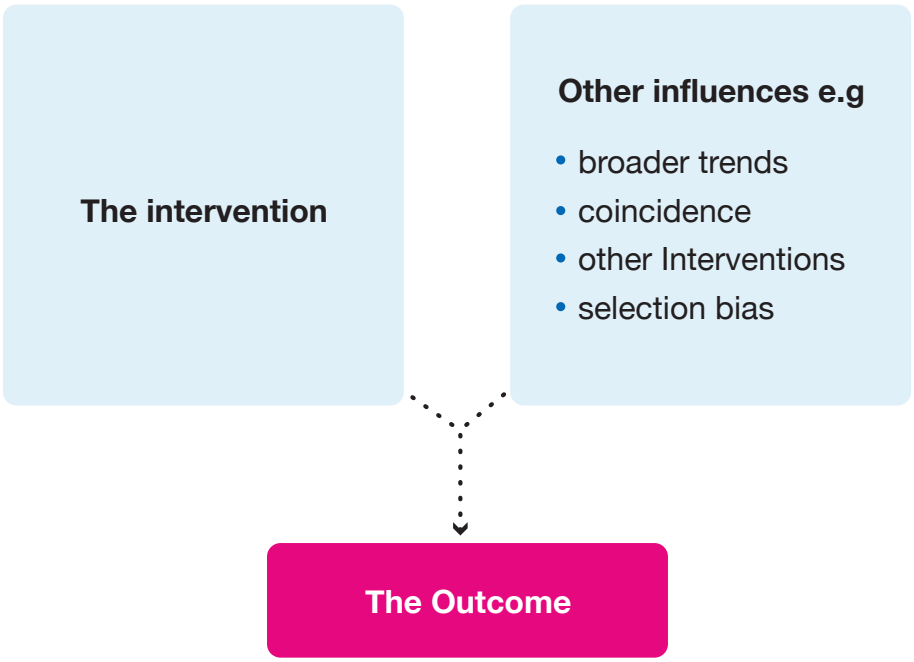
- Are the intervention's goals relevant, in different contexts?
- Can the policy be expected to work in other contexts?
- Is the intervention sustainable from financial, economic, social and environmental perspectives?
- What has been learned about how to intervene in this intervention space that can be transferred to other initiatives and future appraisals?

2.2.4. Understanding impact evaluation approaches

In choosing the most suitable evaluation approach(es), an evaluator must consider the type of evaluation and question(s) to be answered, alongside an understanding of the intervention itself, the context in which it is being implemented and information/ data available.

Impact evaluations can be particularly challenging to design and implement. Impact evaluations aim to assess what changes have occurred and the scale of those changes. They also assess the extent to which the changes can be attributed to the intervention, over and above what would have happened had the intervention not taken place. This is complicated because there will be other influences that must be understood to claim that the intervention has had an effect (see Figure 2.3 below).

Figure 2.3: What influences the outcome?[27]



No one evaluation approach can appropriately evaluate all types of intervention – each design has its advantages and disadvantages – and often approaches may need to be combined.

Selecting the impact evaluation approach is an early decision that will influence all subsequent steps. The remainder of this chapter will focus on this choice.

Figure 2.4 below provides a decision-tree on the most appropriate impact evaluation approach.

27 Stern, E., Stame, N., Mayne, J., Forss, K., Davies, R. and Befani, B. (2012) *Broadening the Range of Designs and Methods for Impact Evaluations*. Department for International Development. Available at: https://www.oecd.org/derec/50399683.pdf (Accessed: April 2026).

Figure 2.4: Selecting the approach for impact evaluation, based on the evaluation questions to be answered[28]

To what extent have outcomes been achieved?

Yes

The intervention:
- is single purpose
- is local
- expects outcomes in 1–3 years
- is introduced to a stable context
- has few or no alternative explanations for change in outcomes

Yes

Before-and-after outcome study

No

Have outcomes been achieved and was it the intervention that led to the impact?

Yes

The intervention:
- can be isolated from other interventions
- has a single outcome
- will demonstrate short-term outcomes
- is based on robust hypothesis
- remains static
- expects small to medium-sized outcomes (change)

Yes

Experimental approach

- the intervention targets individuals
- no interaction between control and experimental (treated) group
- stable population base
- the experimental group is not atypical of the population

No

- the intervention targets groups or areas
- movement/interaction between control and experimental groups is expected
- the population base is subject to change
- the experimental group is not atypical
- the target group is very small

Yes

Experimental methods

Yes

Quasi-experimental methods

No

What changes have taken place / outcomes occurre and why?

Yes

The intervention:
- is developmental or exploratory
- sits in a dynamic environment
- has unclear or soft outcomes
- may have complex causal pathways
- may vary in implementation
- is implemented in a context that is likely to matter
- has a medium to long-term time scale

Yes

Theory-based approach

The evaluation intends to:
- test implementation theory
- explore interaction between multiple outcomes
- work with stakeholders
- seek generalizable conclusions

The intervention:
- is highly complex / multiple interventions / multiple sectors
- has indirect effects
- is long term

Yes

Theory-based methods

No

The evaluation intends to:
- explore programme logic
- explore underlying mechanisms

The intervention:
- will be highly context-dependent
- has limited focus and geographic coverage

Yes

Realist evaluation method

Combined design

28 Hills, D. and Junge, K. (2010) *Guidance for Transport Impact Evaluations: Choosing an Evaluation Approach to Achieve Better Attribution*. The Tavistock Institute in consultation with AECOM. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/525806/transport-impact-evaluations.pdf (Accessed: April 2026).

Experimental and quasi-experimental approaches to impact evaluation

Experimental and quasi-experimental approaches are designed to achieve a robust estimate of the average impact of an intervention. There are a number of experimental and quasi-experimental methods, summarised in section 3.5 for further detail, in Annex A.

Experimental and quasi-experimental approaches infer the impact of an intervention through statistical comparison to a group or time period unaffected by the intervention. This unaffected group acts as a proxy for what would have happened to the affected group in the absence of the intervention and is commonly called the counterfactual.

When measuring the counter-factual data it is essential that:

- It is of sufficient quality and quantity to support the analysis.
- The counterfactual is genuinely comparable to the intervention group.
- The intervention effect is sufficiently large to be distinguished from expected 'noise' in the data.

To meet these requirements often requires building the evaluation design into the intervention design through creating control groups and ensuring data is collected in both the intervention and the control group. Table 2.3 below shows where an experimental approach is more feasible.

Table 2.3: Choosing an experimental or quasi-experimental approach to impact evaluation

	More feasible if:	Less feasible if:
Nature of intervention	• The intervention is discrete (can be disentangled from other programme interventions) and is stable. • The system the intervention is being applied to is relatively stable and unchanging.	• The intervention is difficult to define or disentangle from other programme interventions or local context. • The intervention adapts over time. • The intervention is being applied to a complex and emergent system (see Supplementary Guidance on Evaluating Complexity for more detail).
Nature of impact	• There is a direct, linear relationship between the expected outcome and the intervention. • A large effect relative to other changes is expected. • The effect is realised within a short time period (and does not vanish immediately thereafter).	• There is a complex[29] or distant relationship between the expected outcome and the intervention, with many potential confounding factors. • A small effect is expected. • The effect builds up gradually over an extended time period. • The exact nature of the impact is unknown.
Data availability: what was done where, when, to whom	• The intervention involves a distinct change in practice with respect to identifiable participants (individuals, groups, institutions or areas). • Data is available on the participants of the intervention. • Data is available on precise time periods. • Data to support the evaluation collected before and during the intervention. • Data can be collected from samples of sufficient size.	• The intervention involves a consolidation of existing good practice or is poorly differentiated between participants. • Data is only available as coarsely aggregated totals. • There is uncertainty over timing of implementation (requires aggregation over time). • Data to support evaluation is not sought until the policy is already established, or is unavailable for non-participants. • Sample sizes are too small.

29 M Treasury (2020) *Magenta Book Supplementary Guidance: Handling Complexity in Evaluation*. London: Crown Copyright. Available at: https://www.gov.uk/government/publications/the-magenta-book (Accessed: April 2026).

	More feasible if:	Less feasible if:
Potential Comparison groups	• Intervention is built into policy design, and so comparison groups are allocated and data collected from both. • There is a phased start. • Random allocation is possible. • Other objective allocation is possible, for example, using a cut-off score. • 'Natural' comparison groups are available.	• Intervention is not built into policy design, or data is available only for the pilot areas themselves. • There is a simultaneous launch nationwide. • There is subjective allocation, so the comparison and target group are different from the outset. • No equivalent comparison group is possible (like for a major infrastructure scheme).

Experimental and quasi-experimental approaches tend to be most suitable where:

- There is a focus on accountability. These approaches typically produce the most defendable quantitative evidence on impact.
- The expected outcomes are known and measurable.
- The intervention affects large numbers of people/groups, but not the whole population.
- The intervention does not involve a number of different activities or a varied implementation.
- The intervention is expected to work in the same way for different groups. Typically, experimental approaches are not strong in identifying different impacts on particular sub-groups.

Experimental and quasi-experimental approaches are selected with the primary aim of assessing the net impact of an intervention. But this in itself will not produce insights about how any measured change comes about, or whether the same outcome would occur if the intervention is tried in another context or at a different scale.[30] Combining experimental/quasi-experimental approaches with theory-based approaches or supplementing with process evaluation evidence can provide this often-essential insight.

30 Terms often used are internal validity, which refers to the robustness of the estimate of the particular policy evaluated; and external validity, which refers to whether the estimate is a valid representation of what might happen under non-experimental conditions when the intervention is scaled up or rolled out. External validity can be developed in two ways, through testing a broad number of contexts in a pilot and/or understanding how something has worked and inferring if the evidence would therefore suggest it could work elsewhere.

Theory-based approach to impact evaluation

Theory-based impact evaluations draw conclusions about an intervention's impact through rigorous testing of whether the causal chains thought to bring about change are supported by sufficiently strong evidence and that alternative explanations can be ruled out. Theory-based evaluation is explicitly concerned with both the extent of the change and why change occurs; it tries to get inside the black box of what happens between inputs and outcomes, and how that is affected by wider contexts.

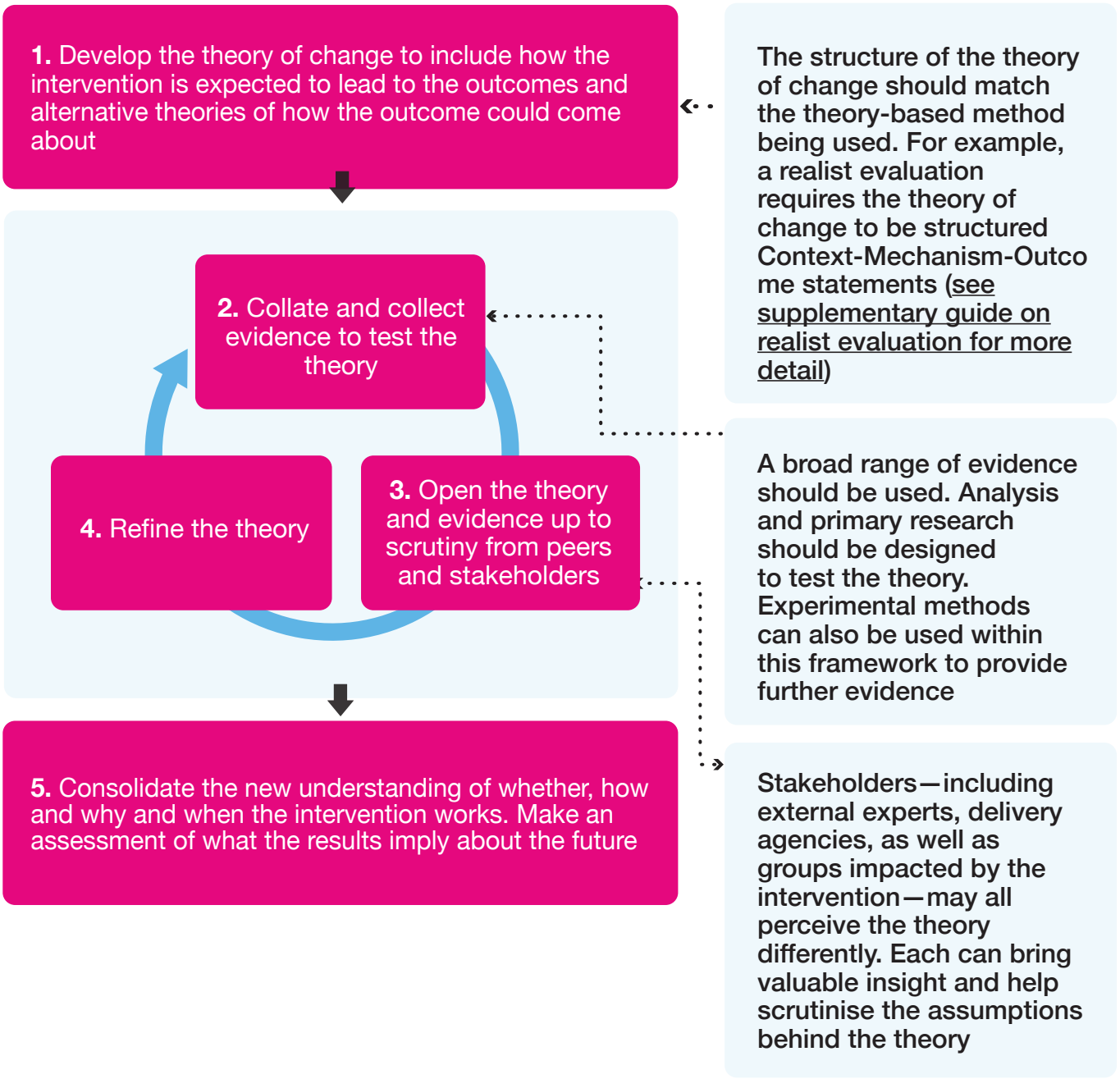
Theory-based evaluations are centred on a well-defined Theory of Change, which includes theories about alternative explanations for the outcomes. Once a theory is established, the theory is tested through multiple evidence sources. A Test and Learn approach can contribute substantially by building and refining the Theory of Change in advance, and by identifying the most critical contribution or causal claims to be examined in theory-based impact evaluations.

When conducting a theory-based approach, it is important to keep an open mind about what evidence can be used. External literature, expert opinions, public statements, mixed-method research with a range of stakeholders and modelling private cost/benefit trade-offs can all be used. There can be efficiency in data collection, by focusing effort on collecting evidence that is impact-orientated.

Rigour in theory-based methods comes from:

- Coherence of the theories.
- Evidence that is specific enough to test the theories.
- Triangulation of multiple sources.
- Ruling out of alternative causes in order to claim impact.
- Critical reflection and opening up to peer review and external scrutiny.

Figure 2.5: Process followed using a theory-based approach



Theory-based impact evaluation methods are particularly suited where one or more of the characteristics below are true:

- There is a complicated policy landscape with a combination of interventions.
- The intervention is designed to make a change in a complex system or where there is adaptive management/changing of an intervention.[31]
- Outcomes are emergent and cannot be predicted at the outset.
- There is no ability to develop a suitable counterfactual.
- There is a requirement to understand if the same results would be achieved in a different place or context.

Typically, these approaches do not produce a single numeric estimate of impact, although quantification of impact may be possible, given the appropriate methods chosen (see Chapter 3).

Participatory and other approaches

Participatory and other approaches to evaluation can provide more tools for use in the evaluation of complexity. A common approach is developmental evaluation (Patton, 2006[32]), which is premised on dealing with uncertain or emerging outcomes and is intended to be part of a continuous loop of programme development and iteration.

Developmental evaluation approaches are often applied to innovation contexts, where it may help to frame thinking, surface issues and provide rapid feedback on the implementation of interventions.

There are several other emerging alternate approaches for handling complexity. See the Magenta Book Supplementary Guidance on Evaluating Complexity for further information.

Direct observation of impact

In some very simple cases, the way in which the intervention creates change may be sufficiently transparent that the impact can be observed directly, or through a process evaluation, without the need to take account of other influences. For example, with a project to supply water to a village in a developing country, any observed decreases in the average time household members spend collecting water can be attributed to the project without the need for a comparison group.[33] This approach should only be used when there is confidence that if the intervention had not taken place nothing else would have changed.

31 HM Treasury (2020) *Magenta Book Supplementary Guidance: Handling Complexity in Evaluation*. London: Crown Copyright. Available at: https://www.gov.uk/government/publications/the-magenta-book (Accessed: April 2026).
32 Patton, M.Q. (2006) 'Evaluation for the Way We Work', *The Nonprofit Quarterly*, 13(1), pp. 28–33. Available at: https://nonprofitquarterly.org/evaluation-for-the-way-we-work/ (Accessed: April 2026).
33 White, H. (2009) *Some Reflections on Current Debates in Impact Evaluation*. New Delhi: International Initiative for Impact Evaluation (3ie). Available at: https://www.3ieimpact.org/evidence-hub/publications/working-papers/some-reflections-current-debates-impact-evaluation (Accessed: April 2026).

2.3 Designing an evaluation

Designing an evaluation involves taking the conclusions of the scoping stage and agreeing the evaluation design and how the evaluation will be conducted. It involves making pragmatic decisions on evidence priorities considering timelines, resources and feasibility of methods. An evaluation design should be guided by the principles outlined in Chapter 1, such as usefulness, credibility, robustness and proportionality.

Designing an evaluation is an iterative process; decisions on methods will impact on earlier decisions on which questions can and cannot be answered. With each iteration the design gets more detailed. An initial evaluation design should give enough detail to make decisions about which evaluation questions are feasible to answer, and which methods would be needed to do this. A formal evaluation feasibility assessment is a common interim design stage, sometimes called an 'evaluability assessment'.

A final evaluation design is specific about how the evaluation methods, research and analysis will answer the evaluation questions, test the intervention logic and report on a timeline to support the identified decision-points.

These details are usually set out in an evaluation 'plan' or 'strategy', which cover information such as:

- The objectives of the intervention.
- The Theory of Change.
- A plan for any literature review/ synthesis required.
- The evaluation questions and by when and from whom the evidence is needed.
- How outcomes of interest will be identified and measured.
- How unintended outcomes will be detected.
- How wider contextual factors will be accounted for.
- Detail on the methods to be used/how they answer the evaluation questions.
- Detail on the data required, how it will be collected and by whom.
- Resource (funding, staff, skills) requirements.
- An overview of how and when the evaluation evidence will be disseminated and used to inform policymaking.

As the evaluation plan develops, decisions will need to be made on the methods and data collection required to answer the evaluation questions. Evaluation methods are covered in Chapter 3, and data collection in Chapter 4.

Sources of further information

Stern, E., Stame, N., Mayne, J., Forss, K., Davies, R. and Befani, B. (2012) *Broadening the Range of Designs and Methods for Impact Evaluations*. Department for International Development. Available at: https://www.oecd.org/derec/50399683.pdf (Accessed: March 2026).

White, H. (2009) *Some Reflections on Current Debates in Impact Evaluation*. New Delhi: International Initiative for Impact Evaluation (3ie). Available at: https://www.3ieimpact.org/evidence-hub/publications/working-papers/some-reflections-current-debates-impact-evaluation (Accessed: April 2026).

2.3.1. Considerations for evaluating a group of projects or programmes

It is common to need to evaluate an intervention that consists of a group of related but separate projects. There are two main scenarios:

1. Where an overall intervention has broad coverage, but the detailed implementation is delegated to a lower level (within set parameters). This might be done geographically[34] (such as in the Well North programme) or by industry sector (such as the Industrial Strategy).
2. A scenario often described as an 'innovation fund', where competitive proposals are invited to deliver an intervention, again within broad parameters, with an emphasis on innovation.

Such scenarios give rise to some particular evaluation issues. An important question is to decide on the balance of importance between evaluation of the overall intervention, and of the specific projects.

Typically, in the first scenario, there will be a strong interest in evaluating the overall net impact of the intervention, and the process of allowing a degree of autonomy. There is also likely to be interest in what can be learned from the individual implementations, but not necessarily in evaluating each one separately. Depending on the specific circumstances, techniques, such as Qualitative Comparative Analysis (see Annex A: Analytical evaluation methods, section 1.1), could be useful in understanding in broad terms how the more successful projects varied from those that were less so. It may be useful to select some more detailed case studies, based on an assessment of, for example, which ones are particularly successful, or which seem particularly innovative.

34 See section 2.3.2 *Considerations when evaluating place-based interventions*.

In the second scenario, where the purpose of the overall intervention is likely to be to identify promising approaches, there will be a much stronger emphasis on evaluating individual projects. Typically, a net impact estimate for each will be desirable, and as far as possible, an assessment of the potential generalisability of the intervention. It must be borne in mind that this will only be a broad assessment. To estimate reliably what might happen if any particular intervention were delivered more widely would require further testing. At the overall level, a key question is likely to be the extent to which the intervention design encouraged useful innovation.

Common to both scenarios is the need to agree and comply with common measures across all projects (as far as possible). (The qualification 'as far as possible' reflects the fact that different projects within the same programme might be seeking to influence different outcomes, and use different processes, in which case it will not be possible to have a single set of measures.) It can be particularly advantageous if some of these measures can be derived from administrative systems, ensuring comparability. If survey measures are needed, using a common research instrument across all projects is necessary. Particular care needs to be taken when requiring individual projects to collect specific data. As well as setting out clear and detailed instructions on what to collect, the importance of keeping to these instructions will need to be explained.

Some degree of compliance checking may be advisable.

In such programmes, it is likely that interventions will be evolving during the programme, and consideration should be given to a closer relationship between the projects and the evaluation team, such as in Developmental Evaluation (see Participatory Approaches section).

2.3.2. Considerations when evaluating place-based interventions

Place-based interventions aim to achieve outcomes for a place, a defined and recognised geographical area. They can be single interventions or, more commonly, multiple interventions that have a common aim. The distinguishing characteristic is that the treatment unit will be defined geographically; it could be a local authority, a town or city centre, a neighbourhood, or an area designated for regeneration. The aims of place-based interventions may include, for example, economic regeneration, increased employment, reduced crime or increased public involvement and satisfaction.

Place-based business cases are intended to sit above project business cases to improve the assessment of how different types of projects, such as housing, public realm improvements and transport, complement each other in achieving objectives for a place. In a similar way, evaluation of place-based interventions should build on evaluations of individual projects to assess what place-based outcomes they achieve collectively, which may be greater than the sum of their individual outcomes.

Where multiple projects interact at a place, it may be challenging for evaluation to isolate and attribute their individual impacts (since each project would be, in effect, a confounding factor in evaluating the impact of the others). Consequently, the strongest evidence for the impact of investment might be obtained through a place-based evaluation approach that considers the collective impacts of complementary projects.[35, 36]

Evaluation of place-based interventions raises some distinct considerations.

- **Place-based theory of change.** It is useful to develop a place-based theory of change that can build on the theories of change for individual projects and describe how they are expected to interact to deliver outcomes for the place. Where projects are complementary, these outcomes may exceed the sum of the expected outcomes for individual projects. The place-based theory of change can be informed by past evidence. For example, a review of the transformational impacts of transport investments conducted for the Department for Transport suggested some conditions where transport investment could enable more housing development in an area.[37] Such conditions can be captured in a theory of change, identifying mechanisms to be tested in the evaluation.
- **Spatial definition.** To evaluate place-based interventions, it is necessary to define the area receiving treatments from them. This definition may reflect administrative boundaries or the locations which are targeted for treatment. It may need to be adjusted to reflect the units at which spatial data are available. Having defined the treatment area, measurement of two types of spatial effect may be considered:
    - Spillovers – wider outcomes, either positive or negative, that occur in neighbouring areas that are not targeted by the intervention.
    - Displacement – where an improved outcome in the treatment area may result in a corresponding deterioration somewhere else.

  Understanding these spatial effects will improve the evaluation's assessment of impact and value for money.
- **The selection problem.** Places are not selected for investment randomly but because of distinct characteristics – either a relative strength that interventions will seek to exploit or a relative weakness that they will seek to mitigate. Consequently, it may not always be possible to identify comparison areas that are sufficiently well matched to be used for counterfactual analysis. A solution to this problem is to use a synthetic control area, which is defined statistically, using available data, to mimic the characteristics of the treatment area (see Table 3.3). Where a robust counterfactual is not achievable, theory-based impact evaluation approaches may be used to examine the causal mechanisms that are expected to deliver outcomes. This approach was used in the Ministry for Housing, Communities and local government's evaluation of the UK Shared Prosperity Fund.[38]

35 The What Works Centre for Economic Growth (2022) *How to evaluate*.
Available at: https://whatworksgrowth.org/resource-library/how-to-evaluate (Accessed: April 2026).
36 Youth Endowment Fund (2023) *Evaluating place-based approaches: a review of methods used*.
Available at: Evaluating-place-based-approaches.pdf (Accessed: April 2026).
37 Department for Transport (2023) *Transformational impacts of transport investment: QCA technical report*.
Available at: https://assets.publishing.service.gov.uk/media/6437f375773a8a0013ab2bd0/transformative-impacts-of-transport-investment-qca-technical-report.pdf (Accessed: April 2026).
38 UK Government (2025) *UK Shared Prosperity Fund (UKSPF) Place-Level Methodology Report*.
Available at: https://www.gov.uk/government/publications/uk-shared-prosperity-fund-ukspf-place-level-methodology-report
(Accessed: April 2026).

- **Community engagement and participation.** For some place-based interventions, there will be significant interest and involvement from local communities and stakeholders in how they are delivered and what outcomes they produce. Accordingly, evaluation of such interventions may gain understanding and enhance local credibility by use of participatory approaches that build in engagement with these communities and stakeholders in their design, delivery and the communication of their findings.
- **Meta evaluation.** It can be challenging to draw robust conclusions about the impact of place-based interventions where only one place is examined. Stronger evidence can be obtained from meta-evaluation approaches which use evidence from multiple places that receive similar interventions, ideally matched with multiple comparison areas (see section 3.7). Looking at evidence from multiple areas may make it possible to explore which mixes of complementary projects are most effective in delivering place-based objectives, providing generalisable findings that cannot be obtained from single interventions.
- **Evidence lags and retrospective evaluation.** A final consideration for place-based interventions is that they often entail a long wait for evaluation evidence, due to the time required for delivery (especially for infrastructure schemes), the accumulation of outcomes and the provision of spatial data. While waiting for evidence from current schemes, evidence gaps may be addressed through retrospective evaluations of places where similar mixes of projects have been delivered in the past. Retrospective evaluation is often feasible for place-based interventions due to the availability and quality of administrative data.

Attention to these considerations in the design of evaluations of place-based interventions can strengthen the evidence that will be available to inform future place-based business cases. However, these types of interventions are challenging to evaluate robustly, so there may remain limitations on what can be concluded that is generalisable.

2.3.3. Registering an evaluation

All planned, live and completed government evaluations from 1st April 2024 onwards must be registered on the Government Evaluation Registry. This applies to planned evaluations, live evaluations and evaluation reports produced by all central government departments. It also includes all evaluations of government major projects, including where these are delivered and/or evaluated by external organisations.

Registering an evaluation on the Evaluation Registry requires only minimal detail on the planned research. In addition to this mandatory requirement, however, all evaluation planning documents – including evaluation plans, protocols and statistical analysis plans – should be completed before the evaluation begins, time-stamped, and preserved within the department. These documents should be made available to relevant stakeholders throughout the evaluation life cycle.

Once an evaluation has been designed, it is also good practice to publish the evaluation plan or strategy in a process known as pre-registration. Pre-registration means documenting key elements of the evaluation such as its objectives, research questions, design, data collection procedures and analytical approach, before any outcome data is collected or analysed. For quantitative evaluations, pre-registration should also include power analyses and sample size calculations (see section 3.5.1 below). Once a document is produced, it can then be

published in order to create a transparent record of what was intended, which can later be compared with what was done in practice.

Pre-registration supports the principles outlined in Chapter 1. It helps ensure evaluations are **useful**, by clarifying in advance how evidence will be produced to meet the needs of stakeholders. It strengthens **credibility**, by making the intended evaluation design transparent and reducing the risk that unfavourable or partial results are withheld. It contributes to **robustness**, by encouraging well-specified methods and analytical choices that limit the scope for post-hoc adjustments and help ensure the evaluation is capable of providing reliable evidence.

Pre-registration does not prevent changes being made to an evaluation design, but it requires that amendments to evaluation plans are documented and explained. The purpose of pre-registration is to prevent researchers from 'fishing' for significant or positive findings by arbitrarily adjusting the way that data is collected or analysed. This does not mean that researchers must always conduct their analyses in the way that was pre-specified, but rather that – when changes are made – evaluators clearly explain how the analysis was planned at the outset, and how it changed in the production of the final report. This provides transparency about why changes were made while giving evaluators the flexibility to adapt when circumstances demand it.

Pre-registration should be applied in a **proportionate** way. Evaluation teams should adopt more detailed forms of pre-registration in line with the scale and importance of the project. For smaller or lower-risk evaluations, a short study registration may be sufficient. For larger, more resource-intensive or higher stakes evaluations, pre-registration should normally include a full evaluation protocol and a pre-registered statistical analysis plan.

Further detail on pre-registration is provided in the Transparency in Government Evaluation Research (TIGER) annex.

HM Treasury

Chapter Three

Evaluation methods

3

Summary

There are a wide range of evaluation methods. This chapter summarises the main analytical methods used in government and their key pros and cons. Further details are contained in Annex A.

First, generic quantitative and qualitative research methods that are commonly used for both process and impact evaluations are summarised.

Next, evaluation methods are set out and are grouped into:

- Theory-based impact evaluation methods.
- Experimental and quasi-experimental impact evaluation methods.
- Value for money methods.

Some guidance on how to select appropriate methods for addressing causal questions is provided. Selection of methods should always be considered carefully and in context. Stakeholder involvement is strongly recommended as the choice of method affects the questions that can be answered (choice of method may depend on which assumptions are most plausible, and stakeholders may have a better perspective than government analysts on this).

Methods for the synthesis of evaluation evidence are also summarised. These methods are used to combine findings from a range of studies into a common understanding of the impact and/or delivery of an intervention.

3.1 Introduction

This chapter covers the considerations required to complete the evaluation design; that is, the selection of the analytical evaluation methods that best answer the evaluation questions (in contrast to the data collection evaluation methods discussed in Chapter 4). Alongside the evaluation questions, three further considerations are important: the proportionality of the effort and resources required to answer them; the availability of data; and the timing of decision points.

As stated in the previous chapter, iteration will be required to refine what is possible and practical. This might require a change to the scope, timing or resources or even modification of the evaluation questions. Different methods will have different resource implications. Any deprioritising of evaluation questions should be communicated clearly and the solution agreed with stakeholders.

Some methods will be able to answer more questions than others and some will be more cost-effective and easier to implement; all have their pros and cons.

Detail on each of the methods summarised in this chapter can be found in Annex A. Methods that are particularly appropriate for evaluations in complex situations can be found in the Supplementary Guidance, 'Evaluating Complexity'. Data collection methods are covered in Chapter 4.

3.2 Choosing the appropriate methods

The selection and implementation of analytical methods for evaluation should be informed by the Theory of Change and the uncertainties and assumptions that it identifies.

It is important to note that various stakeholders may have different needs and prioritise different outcomes. These outcomes should be reflected in the Theory of Change and stakeholders should be engaged in the selection of the evaluation methods in order to agree priorities, manage expectations and accept the implications.

There should be alignment between the evaluation questions, the evaluation approach and the methods used. In practice, most evaluation designs involve mixed methods, combining both qualitative and quantitative methods to answer the impact, process and value for money questions (Chapter 2). No individual method can provide answers to all evaluation questions. If time and resources are limited, questions and methods will have to be prioritised and trade-offs between methods will be needed.[39] Iteration will be necessary.

It is important to consider both the feasibility and appropriateness of a proposed method. The cost, timing, respondent burden, ethics, likely response rate and the potential effect the data collect itself might have on the intervention should all be considered (see Chapter 4 on data collection).

The populations to be studied and the evaluation questions to be answered should be clear from the scoping stage. Clear definitions should be developed around the individuals/groups, places and time periods associated with the evaluation questions to ensure the most appropriate methods are identified. Clear definitions can also help clarify the extent to which the findings can be generalised.

For many evaluation methods, the option exists to integrate AI into the workflow. If using AI, evaluators must weigh the analytical power of these tools against the challenges they can introduce, such as the 'black box' nature of some models which may obscure the link between evidence and conclusions. The appropriateness of using AI must therefore be judged on a case-by-case basis, ensuring the chosen approach is proportionate, quality-assured, and verifiable, and transparently communicated to stakeholders. (See section 5.6.)

The timing of data collection needs to be carefully considered in line with when events are likely to occur and the decision points identified during the scoping phase (see Table 2.1).

39 Bond (2016) *Choosing Appropriate Evaluation Methods Tool*. Available at: https://www.bond.org.uk/resources/evaluation-methods-tool (Accessed: April 2026).

3.3 Research methods used in both process and impact evaluations

The research methods summarised in this section are commonly used in process and impact evaluations (and are widely used in non-evaluative research). When used in impact evaluation, consistent data must also be collected for any control/ comparison group.

Research methods commonly used for evaluation are outlined in Table 3.1 below, with more detail in Annex A. The data collection aspects of these methods are covered in Chapter 4. Detailed guidance on process evaluation and its methods can be found elsewhere.[40] Guidance on quality is provided in the Supplementary Guidance, 'Quality in Qualitative Evaluation'.

40 Moore, G., Audrey, S., Barker, M., Bond, L., Bonell, C., Hardeman, W., Moore, L., O'Cathain, A., Tinati, T., Wight, D. and Baird, J. (2015) *Process Evaluation of Complex Interventions*.
Available at: https://www.ukri.org/publications/process-evaluation-of-complex-interventions/ (Accessed: April 2026).

Table 3.1: Quick guide to research methods commonly used for evaluation

Evaluation methods	Description	Pros and cons
Interviews	Interviews enable in-depth exploration of the intervention with individual participants.	Can be used to elicit the views of individuals involved in an intervention. Can be used to collected in-depth insight about an intervention and shed light on patterns emerging in other pieces of evidence collected (such as quantitative monitoring data). Can be resource intensive; requires time to conduct and analyse; does not provide numerical estimates; there may be risk of bias in the views collected.
Focus groups	Focus groups are useful to elicit views from a group of people rather than an individual.	Can be used to elicit views from a group of people. Can be used to collected in-depth insight about an intervention and shed light on patterns emerging in other pieces of evidence collected (such as quantitative monitoring data). Can be resource intensive; requires time to conduct and analyse; does not provide numerical estimates; there may be risk of bias in the views collected.
Case studies	In-depth investigation of a person, group or event within its real-world context. Subjects are often purposively selected because they are unusual and reveal information. Often uses multiple sources of evidence and data collection methods. Can be descriptive, exploratory or explanatory.	Can capture real-life situations in depth and detail and help understand complex phenomena. Works well in combination with or supplementing other methods, such as surveys. Can be helpful for communicating to stakeholders what interventions have worked for particular organisations in certain contexts. It is difficult to generalise findings to different contexts, situations or phenomena.

Evaluation methods	Description	Pros and cons
Surveys	Commonly used to collect data from a number of individuals, such as beneficiaries, or a large organisation with numerous members of staff. They can be administered face-to-face, by post, online, by telephone or as a handout.	An effective method of obtaining information from a large number participants. Provides data suitable for statistical analysis that, if properly designed and conducted, can be generalised to the whole population of interest. Less useful for providing in-depth insight into an intervention. There can be response-rate issues that decrease the quality of its findings.
Output or performance monitoring	Continuous measurement and performance review of an intervention. Monitoring plans are developed based on the Theory of Change to allow the tracking of the inputs, outputs and outcomes of an intervention. To minimise errors in the data collection and burden on staff it is advisable to design monitoring in collaboration with those who will collect data.	Can provide a relatively low-cost and rapid method to identify if an intervention is being delivered and creating outputs as intended. Links well to benefits management. Can feel onerous for both participants and the staff (usually delivering an intervention) who collect it.
Observational studies (including ethnography)	These involve observing and noting behaviour of participants (including intervention delivery staff). Often supplemented with interviews of individuals within their usual environments to determine the impact that an intervention has had on an individual, day to day life. Can support other methods by ensuring that other data collected are understood within the same context and are used to build theories relevant to this context.	Can allow for a deeper understanding of individual experience of an intervention. Observation may help improve accuracy of other data by reducing bias arising from self-reporting by participants. However, participants may still act differently if they know they are being observed (the 'Hawthorn effect'), which can affect the accuracy of the data. Resource-intensive and may have ethical implications, practical barriers and issues with generalisability.

3.4 Theory-based impact evaluation methods

Theory-based methods can be used to investigate net impacts by exploring the causal chains thought to bring about change by an intervention. However, they do not provide precise estimates of effect sizes. Theory-based evaluation is explicitly concerned with both the extent of the change and why the change occurs. In addition, it often considers the context at the same time that the intervention is being implemented.

Theory-based methods tend to be particularly suited for the evaluation of complex interventions or simple interventions in complex environments. In these situations, where determining the effect size can often be difficult, theory-based methods can confirm whether an intervention had an effect in the desired direction. For many of these methods, the aim is not to provide definitive evidence that the entirety of any measured change can be attributed to the intervention. Rather, they aim to explore whether the intervention definitively contributed to the measured change. They can also explain why an intervention worked, or not, and inform translation to other populations, places or time periods.

All evaluation methods can be considered and used as part of a theory-based approach (particularly research methods that aim to answer process evaluation questions), but those described below are particularly associated with this approach.

As an example, 'realist evaluation' seeks to identify the often psychological mechanisms that change human behaviour as a result of an intervention, taking into account the context within which the intervention occurs. Realist evaluation typically asks 'what works, for whom, in what respects, how, and under what circumstances?'. It develops and tests a set of hypotheses (or theories) about the factors or processes that explain why an intervention has had a particular result (called a mechanism), and what effect the context of an intervention has on these mechanisms. A mechanism can be defined as capturing 'people's reasoning and their choices when faced with a policy measure'. This method is described in detail in the Supplementary Guidance, 'Realist Evaluation'.

Simulation modelling is not formally considered a theory-based method but is included here as it can be used quantitatively to represent complex scenarios and model the Theory of Change, the impact of an intervention and unobserved outcomes. Types of simulation for dynamic systems include discrete event simulation, system dynamics and agent-based modelling. Model strength depends on quality of the inputs and simulation logic, which often have flaws. Simulation modelling can also be used to generate virtual counterfactuals.

Theory-based methods are specialist and typically require expert advice to help develop thinking and decide on the most appropriate method(s). Useful tools and guidance to inform thinking are available.[41,42,43] Table 3.2 below provides an overview and pros and cons of the most common theory-based methods. More information can be found in Annex A.

41 Bond (2016) *Choosing Appropriate Evaluation Methods Tool*.
Available at: https://www.bond.org.uk/resources/evaluation-methods-tool (Accessed: April 2026).
42 Befani, B. (2016) *Choosing Appropriate Evaluation Methods*. Bond.
Available at: https://www.bond.org.uk/wp-content/uploads/2022/08/caem_narrative_final_14oct16.pdf (Accessed: April 2026).
43 Stern, E. (2015) *Impact Evaluation: A Guide for Commissioners and Managers*. Bond.
Available at: https://www.gov.uk/research-for-development-outputs/impact-evaluation-a-guide-for-commissioners-and-managers (Accessed: April 2026).

Table 3.2: Quick guide to common theory-based impact evaluation methods

Evaluation method	Description	Pros and cons
Realist evaluation	Specific, hypothesised causal 'mechanisms' for an 'outcome' are articulated in 'context' and evidence gathered for each. The 'mechanism' explains why participants may take advantage of an opportunity or not depending on the 'context', and their understanding is key to causal inference.	Refines theory can identify causal mechanisms. Can inform impact if a counterfactual is not feasible. Time consuming, resource intensive and needs subject-matter expertise. Often difficult to communicate/ interpret due to complexity. Does not often provide a quantitative effect size.
Contribution analysis	Step-by-step process used to examine if an intervention has contributed to an observed outcome by exploring a range of evidence for the Theory of Change. It gives an evidenced line of reasoning rather than definitive proof.	The contribution claim depends on the quality of thinking about the attribution problem and Theory of Change. Works on average effects – not to be used if there is large variability in implementation or outcomes.
Process tracing	A structured method examining a single case of change to test whether a hypothesised causal mechanism, such as that proposed by the Theory of Change, explains the outcome.	Can test causal hypotheses post-hoc. Must be used with rigour to prevent inferential errors; alternative explanations must be carefully considered. Support for one causal mechanism may not preclude others.
Bayesian updating[44]	Added to other theory-based methods to more rigorously assess whether evidence supports contribution claims. Probabilities of a small number of contribution claims are estimated prior to observation then tested.	Requires highly skilled facilitation.

44 Bayesian updating is often applied as an analytical enhancement of other methods, although it can also be used as a standalone method. See Handling Complexity in Evaluation Annex for a detailed discussion on its application in complex environments. For recent advancements in integrating qualitative and quantitative inferences via Bayesian logic, see Humphreys, M. and Jacobs, A.M. (2015) 'Mixing Methods: A Bayesian Approach', *American Political Science Review*, 109(4), pp. 653–673. doi: 10.1017/S0003055415000453.

Evaluation method	Description	Pros and cons
Contribution tracing	Participatory mixed method to establish the validity of contribution claims with explicit criteria to guide evaluators in data collection and Bayesian updating to quantify the level of confidence in a claim. Includes a contribution 'trial' with all stakeholders to establish what will prove/disprove the claim.	Efficiently focuses on evidence that can increase confidence in a claim. Minimises confirmation bias using 'critical friends' in a testing phase. Intervention needs time to have detectable effects. Must explore other potential causes. Not for comparing interventions.
Qualitative comparative analysis	Used to compare multiple cases and systematically understand patterns of characteristics associated with desired or undesired outcomes based on qualitative knowledge. Can account for both complex causation (combinations of factors) and 'equifinality' (multiple causes of outcomes).	Can identify groups of causal factors in post-hoc evaluation. Systematically analyses case study evidence. Works best with 10–50 cases. Needs consistent data about how those factors affect outcomes and assessment of which are the more successful across case studies.
Outcome harvesting	Collects evidence of change and then works back to assess contribution to the change. Encourages participation of stakeholders in real-time for ongoing monitoring.	Useful where participation is easy and stakeholders are disparate as it helps provide clarity to all. It is resource intensive.
Most significant change	Participatory method for impact evaluation of complex interventions. Involves collection of significant change stories from the field and systematic selection of the most significant by panels of stakeholders. Interventions are often participatory too.	Useful when it is not possible to predict outcomes or when prioritisation of outcomes cannot be agreed. Builds understanding across stakeholders. Is time consuming and resource intensive and needs robust facilitation.

Note: These all allow attribution of causality, but none give precise estimates of effect sizes.

3.5 Experimental and quasi-experimental impact evaluation methods

Experimental and quasi-experimental evaluation methods are used to measure impact. As introduced in Chapter 2, the core principle of these methods is that there is a 'counterfactual': observed outcomes from a 'control' group that did not receive the intervention, which can be compared to outcomes from the intervention group. The intervention and control groups are either effectively identical (typically, through randomisation in an experimental design) or differ in known ways that can be accounted for analytically (in quasi-experimental designs). The groups consist of participants, which might be individual people, or other units, such as schools, businesses, houses, or spatial areas. Mixing between groups should be minimal to limit bias from 'contamination'.

Collecting and analysing comparable data from the intervention and control enables evaluators to confidently attribute any measured change to the intervention (subject to assumptions specific to the method). These impact evaluation methods are favoured when we need to know the average additional or net change caused by an intervention or how much of the observed outcome(s) could be attributed to the intervention.[45] However, the precise method to be used depends on: whether participants in an intervention can be randomised; the expected effect size; the availability of data; and the availability of potential control groups.

A simple, but effective, way to help plan or review experimental methods is the PICOT[46] framework, which considers the population, intervention, control group, outcome and time period.

Randomised Controlled Trials (RCTs) prospectively and randomly allocate participants to either intervention or control groups. RCTs can robustly evaluate the impact of interventions because they account for both known and unknown factors because allocation to the treatment is random. Measured differences between groups can, therefore, be considered to be the result of the intervention alone. However, the need to ensure rigid implementation of an intervention (high fidelity with the protocol) can reduce its external validity.

There are a number of variations of RCTs for various needs and situations: factorial RCTs independently randomise participants to multiple interventions; cluster RCTs randomise groups of participants rather than individuals; and stepped-wedge RCTs apply an intervention sequentially and at random to groups of participants.

Methods, such as Sequential Multiple Assignment Randomised Trial (SMART) and Multiphase Optimisation Strategy (MOST), have been developed for optimisation of interventions and may be particularly appropriate to develop more potent digital interventions.

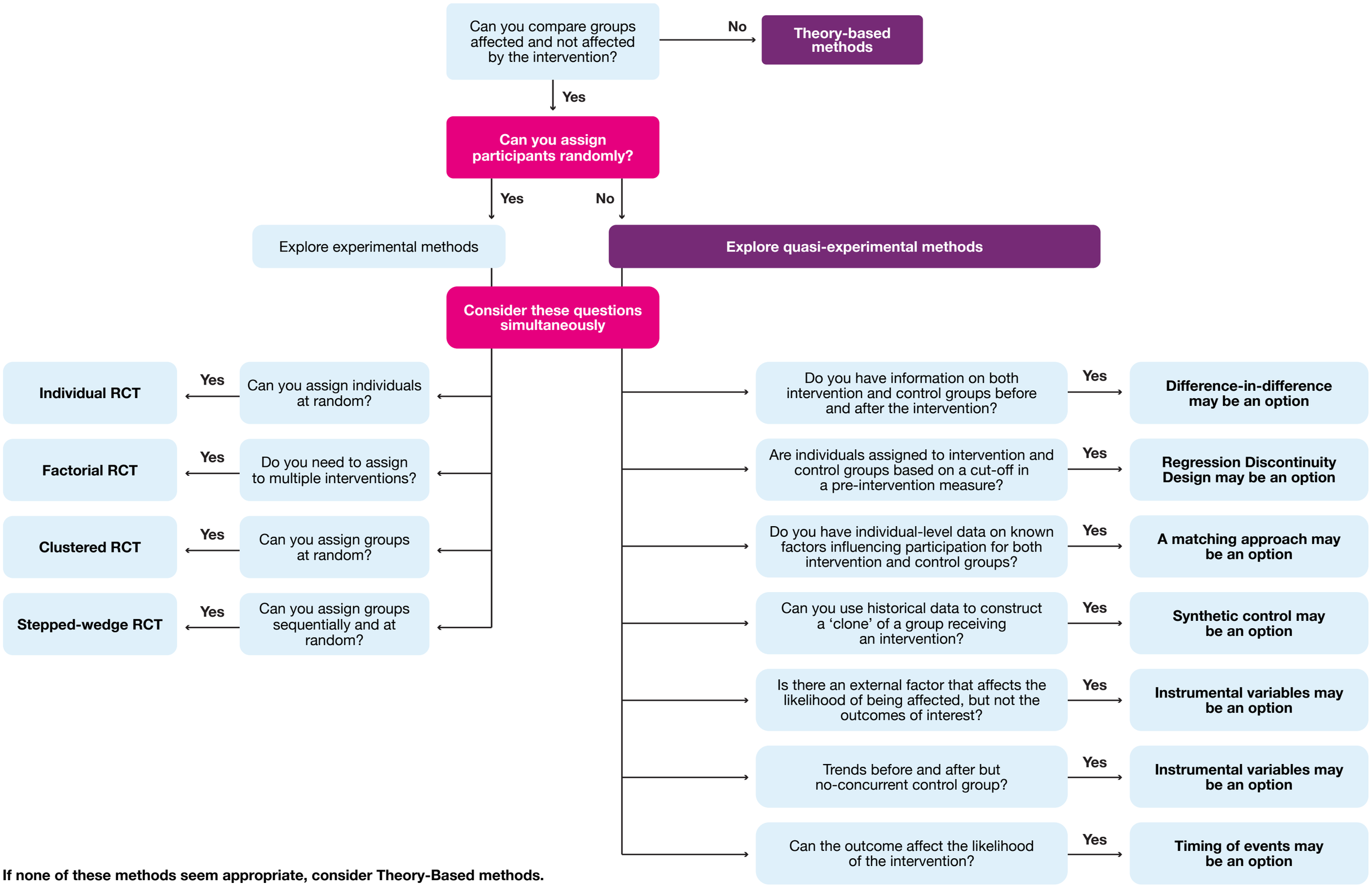
45 Befani, B. (2016) *Choosing Appropriate Evaluation Methods*. Bond.
Available at: https://www.bond.org.uk/wp-content/uploads/2022/08/caem_narrative_final_14oct16.pdf (Accessed: April 2026).
46 Sackett, D., Richardson, W.S., Rosenberg, W. and Haynes, R.B. (1997) *How to Practice and Teach Evidence-Based Medicine*. 2nd edn. London: Churchill Livingstone.

Quasi-experimental methods (see Figure 3.1 below) use a counterfactual, but not one achieved through randomisation. In addition, some methods account for known differences between the intervention and control groups in the analysis.

Selection of a method usually depends on: the nature of the allocation into intervention and control groups (for example, it can be controlled by the evaluator or result from an external shock); the nature of the control group (it can be concurrent or historic); the format of the available data (such as discrete or trends); and the volume of the available data (see Figure 3.1 below).

A set of questions are provided in Figure 3.1 below to help consider which of these experimental and quasi-experimental methods are likely to be most appropriate.

Figure 3.1: Selecting experimental and quasi-experimental methods



If none of these methods seem appropriate, consider Theory-Based methods.

Table 3.3: Quick guide to common experimental and quasi-experimental impact evaluation methods

Evaluation method	Description	Pros and cons
Randomised controlled trials (RCT)	Prospective method to compare the effect of an intervention to what would have happened without the intervention. Randomisation allocates participants to receive the intervention or not.	Prospective nature minimises bias. Randomisation accounts for known and unknown biases. It is not always possible to randomise or to be confident that the control group is unaffected.
Interrupted time series analysis and difference-in-difference	Uses time series data to test for a causal change in the trend of outcomes following intervention. Assumes trends would continue in the absence of the intervention. Difference-in-difference strengthens this through comparison to trends in a control group.	Strong design when randomisation is not possible. The control group minimises assumptions of continuing trends. Difficult if no clear timing of the intervention. Requires time trends before and after with careful consideration of the period to use. Control group can be hard to identify.
Regression discontinuity design	A selection variable is designed into the intervention so that eligibility depends on whether an individual is above or below a set threshold (or inside/outside of a boundary); other known and unknown variables of individuals close to the threshold are expected to be randomly distributed.	Considered causal if observations either side of the threshold/boundary are plausibly random. Needs substantial observations and sensitivity analysis of bandwidth around the threshold. Conclusions may not apply to those further from the threshold/boundary.
Propensity score matching	Statistical technique to create a comparison group that matches the intervention group on all known relevant factors (those which affect both participation and outcomes).	Needs rich data on participants and non-participants to be confident of controlling for all known relevant factors. Bias arises if unobserved characteristics might affect participation and outcomes.

Evaluation method	Description	Pros and cons
Synthetic control methods	Use of historical data to construct a 'synthetic clone' of a group receiving a particular intervention. Differences between the performance of the actual group and its synthetic clone may be used as evidence that the intervention has had an effect. Most commonly applied to interventions applied at an area level.	Offers a relevant comparison when no other comparators exist. Suitable when large volumes of secondary data are already available. Can be used on small sample sizes. Only viable when it can be demonstrated that there was a relationship between the behaviour of the treatment and control groups in the period before the intervention.
Instrumental variables or natural experiments	Needs a factor/instrument that influences selection for an intervention but has no impact on the outcome. Estimates marginal impacts depending on the value of the instrument.	Finding a valid instrument is difficult as many factors will have some association with outcomes. Natural experiments cannot be planned but sometimes can be prepared for.
Timing of events	Estimates the net impact of an intervention by jointly modelling the time when an individual engages in an intervention, and when their outcome changes. Explicitly assumes that timings depend on both observed and unobserved factors.	A substantial advantage is that this allows for unobserved factors that affect selection. Assumes no anticipation effects (outcomes affected by awareness of an intervention before it takes place). Tends to be complex and computationally intensive, and estimations do not always converge.

Note: All of these methods yield net impact estimates, which can be attributed to the intervention, but vary in whether they apply to all participants or a subset. The precision of the estimates will naturally depend on the size of the samples analysed.

3.5.1. Considering statistical power

In any quantitative evaluation, whether experimental or quasi-experimental, evaluators should ensure that they have collected enough data to provide reliable evidence. To determine the appropriate amount of data to collect, evaluators should conduct power calculations, which are analyses that determine the chance that an evaluation can detect the effect of a policy intervention, assuming that it had some minimum effect on the outcome. In other words, a power analysis allows us to assess whether an evaluation's sample size is large enough to detect a meaningful change in key outcomes, if such changes exist.

Impact evaluations which are based on small sample sizes often have insufficient data to draw robust conclusions, particularly for interventions that have small effect sizes.

When the sample size is small, and the effects of an intervention are also small, the study is more likely to produce 'false negative' results, which suggest that a policy had no effect when in reality it did. An evaluation with a small sample size also risks 'false positives' – claiming an effect exists when it does not – which can lead to the adoption of ineffective or even harmful policies.

Evaluators should calibrate the required power level to the context of the policy. Greater power is necessary (and therefore larger sample sizes should be sought) in high-stakes scenarios where failing to detect an effect would be more costly. This includes instances where the intervention involves significant public expenditure, creates non-reversible changes, or carries a recognised risk of adverse consequences.

Power calculations are commonly conducted in general-purpose statistical software, such as R and STATA, but there are other dedicated power analysis software packages such as G*Power. Conducting a power calculation typically requires making assumptions about key aspects of the evaluation, such as the anticipated sample size, the expected effect size, and the amount of variation in the outcome variable of interest.

The results of power analyses should be published. If the evaluation includes the publication of an evaluation protocol or statistical analysis plan (see section 6.6), then power analyses should be included in those documents. If these documents are not published, then the power analysis should be included in an annex to the final evaluation report.

More detailed guidance on conducting power analyses can be found in the Transparency in Government Evaluation Research (TIGER) annex.

3.6 Value for money evaluation methods

Value for money evaluation methods compare benefits to the costs of interventions, including any unintended benefits and costs. It aims to build on and complement process and/or impact evaluation by assessing whether the intervention represents a good use of public resources to achieve the intended outcomes.

The following guiding principles should be considered in designing a value for money (VfM) evaluation:

1. **Embed consideration of VfM within evaluation planning from the outset.** Value for money evaluation is most likely to deliver useful findings when it is planned in alignment with process and impact evaluation. For example, evidence from impact evaluation, such as quantified estimates of an intervention's outcomes, may be used to assess the scale of benefits delivered. Similarly, process evaluation may provide information on cost inputs or on efficiency. How good value is defined should be established at an early stage of the evaluation, before analysis is undertaken.
2. **Link VfM evaluation with appraisal.** Evaluation of the value for money delivered by an intervention should take into account the methods and forecasts from the appraisal stage, which may have used cost-benefit analysis to set out the most important anticipated benefits and costs. Where possible, comparisons should be made between outturn VfM findings and the original expectations outlined within the appraisal and business case. The evaluation should also go beyond the appraisal assumptions to assess any unintended benefits and costs.
3. **Build from empirical data.** As far as possible, VfM evaluation should seek to build from observed data which is relevant to the specific intervention being implemented. This may include outturn data, administrative data or data from primary research. Where possible, VfM evaluation should avoid overreliance on modelling data or assumptions. Some use of forecast data may be necessary where the time period covered by the evaluation does not match the full period over which benefits and costs are expected to be realised, but the evaluation should start with what has been measured empirically.
4. **Present any monetisable findings in context.** VfM evaluation findings should be reported with appropriate context, including acknowledgement of any challenges, limitations or omissions in the analysis (such as any significant unmonetisable or unquantifiable benefits). Where summary measures such as outturn benefit-cost ratios are reported, these should be accompanied by evidence about relevant unmonetisable benefits and costs.

A widely used value for money evaluation method is social cost-benefit analysis, which allows for the comparison of two or more alternative options (interventions).It is designed to produce a robust numerical value which summarises the value for money of the intervention, based on all relevant costs and benefits valued in monetary terms (where proportionate and possible).

Social cost-benefit analysis and its variants listed in Table 3.4 are most appropriate for interventions where the majority of costs and benefits relevant to the intervention can be quantified and monetised. However, the final reported assessment of value for money should also include consideration of any unmonetisable social benefits and social costs which cannot be expressed in monetary terms.

As social cost-benefit analysis is the Green Book's recommended approach for appraisal of shortlisted interventions, using these methods in a VfM evaluation also enables evaluation findings to provide feedback on appraisal assumptions, which can strengthen appraisal practice over time. For example, assessment of whether the observed costs and benefits were within the range originally expected can allow for identification of any systematic forecast errors, which can enable analysts to adjust assumptions for future business cases.

More detail on these common value for money evaluation methods and relevant variants is set out in the table below. Further detail on each method is available in the Magenta Book Annex A on analytical methods for use in evaluation.[47]

47 HM Treasury (2026) *Magenta Book: Annex A – Analytical Methods for Use Within an Evaluation* [online]. Available at: https://www.gov.uk/government/publications/the-magenta-book (Accessed: April 2026).

Table 3.4: Common value for money evaluation methods and approaches

Evaluation methods and approaches	Description	Pros and cons
Social cost-benefit analysis (CBA)	Quantifying and monetising both the costs and benefits of interventions enables comparison between them, even when they have different outcome measures. Can be used to forecast using predicted costs and benefits or used post-intervention in an evaluation using outcome data. It enables a holistic view of intervention options, including broad financial, environmental and social impacts.	Can capture short and long-term impacts systematically. Quality of results relies heavily on available data and ability to monetise the impacts.
Social cost-effectiveness analysis (CEA)	Supports the selection of intervention options that have the same measure of outcome by comparing their costs and effect sizes.	Enables comparison where benefits cannot be converted into monetary units or the cost to do so is prohibitive. Systematically provides comparability of options.
Cost utility analysis (CUA)	A form of CEA which compares the costs of an intervention with its benefits measured in utility-based units (where utility means the satisfaction or benefit that an individual gains). Examples of utility-based units include measures related to wellbeing or quality of life.	Incorporates multiple measures into a single measure of utility, which enables comparability of options aiming to achieve the same outcome. Relies on utility measures capturing all relevant effects of the intervention.

Evaluation methods and approaches	Description	Pros and cons
Social return on investment (SROI)	Measures the social value generated by comparing monetised costs and benefits of an intervention. SROI draws on similar principles to CBA, but is expressed as the social value in pounds per pound invested in the intervention. It may also involve more emphasis on engaging stakeholders to inform the assessment.	Can capture short and long-term impacts systematically. Quality of results relies heavily on available data and ability to monetise impacts.

Further details on value for money assessment methods, including methodologies to monetise benefits, are available in the Green Book.[48]

Some government departments also have supplementary guidance on assessment of benefits specific to certain policy areas, such as the Department for Transport's Transport Analysis Guidance,[49] natural capital approaches for environmental benefits, or the Green Book wellbeing supplementary guidance.

For some interventions, it may not be feasible to quantify and monetise the costs and benefits (for example, where many of the intended outcomes are intangible and hard to measure). In these cases, options such as the Four Es framework or rubric-based approaches may provide a suitable alternative (as set out in Table 3.5). These approaches aim to establish a well-defined theoretical framework which can support a synthesis of the available data and evidence, in order to provide an overall evaluative judgement on the value for money of the intervention.

These approaches can also be used alongside the methods in Table 3.4 to provide a consistent framework for mixed-methods value for money judgements.

While these options offer more flexibility to draw on a range of quantitative and qualitative evidence and data sources, they will generally not produce a numerical value summarising the value for money of the intervention. This can limit the comparability of findings across different interventions.

48 HM Treasury (2026) *The Green Book: Central Government Guidance on Appraisal and Evaluation*. London: Crown Copyright. Available at: https://www.gov.uk/government/publications/the-green-book-appraisal-and-evaluation-in-central-government (Accessed: April 2026).
49 Department for Transport (2025) *Transport Analysis Guidance (TAG)*. Available at: https://www.gov.uk/guidance/transport-analysis-guidance-tag (Accessed: April 2026).

Table 3.5: Additional options for considering and assessing value for money

Evaluation Method and Approaches	Description	Pros and Cons
Four Es framework	This framework sets out four key criteria for assessing value for money, including economy, efficiency, effectiveness and equity. Taken together, the four Es provide a structured approach to assess the overall cost-effectiveness of an intervention. Some frameworks refer to three Es (where equity is not included) or five Es (where cost-effectiveness is considered separately).	Can draw on quantitative and qualitative data to provide a broad assessment of VfM. However, this approach will not generally result in numerical summary measures that can be easily compared across interventions and link with appraisal.
Rubric-based approaches	Rubric-based approaches to assessing VfM involve the development of a set of criteria based on the intended objectives of the intervention, which support an assessment of the intervention's performance against expectations.	Provides a transparent basis for interpretation and synthesis of the available evidence into an overall evaluative judgement on VfM. However, criteria and performance standards are intervention specific. Results therefore generally cannot be compared across interventions or linked with appraisal.

It is worth noting that the methods outlined in Table 3.4 and Table 3.5 can also be used in combination, where a mix of approaches is most suitable for addressing the specific evaluation questions identified in relation to value for money. This may be useful for interventions that have a wide range of both monetisable and non-monetisable benefits.

A combined approach involves establishing an overarching analysis framework for value for money (for example, drawing on the approaches in Table 3.5), and then assessing monetised, quantified and qualitative evidence against this framework. The type of evidence used will depend on what is feasible and proportionate for each part of the framework.

3.7 Synthesis methods

Synthesis methods entail combining findings from a range of studies into a common understanding of the impact and/or delivery of a policy.

Post-evaluation synthesis is used to bring together the outputs of the various methods used in the evaluation in a clear and systematic way to answer the original evaluation questions. The component methods can be quantitative, qualitative or mixed methods, and focus on impact or process evaluation questions.

To synthesise, evaluation findings are integrated around each of the questions (rather than simply reporting the findings from each method), typically called 'triangulation'. Ideally, this provides a consensus of evidence with greater certainty than each of the component parts. However, if there is conflicting evidence, findings should be examined carefully, explanations considered and additional evidence sought where appropriate.

Synthesis methods can also be used prior to initiating the development of an intervention to determine what is already known on a subject.

Meta-evaluation and synthesis are formal synthesis methods designed to bring together findings from a number of different studies on the same subject. Methods of synthesis and meta-analysis bring together existing evaluation evidence in a structured way to produce a coherent narrative. These 'secondary research' methods tend to start with a clearly stated set of objectives and inclusion/exclusion criteria. They then use explicit protocols and pre-defined criteria to assess the quality of source studies, extract key findings from those studies, and analyse the results to provide an integrated account of the source literature.

The main limitation of evaluation synthesis is its reliance on a sufficient body of previously acquired data. The quality of the evidence synthesis depends on the quality of the original studies, and validity depends on which original studies are included. Hence, the importance of inclusion criteria and quality assessment in the process.

As an example, systematic reviews and rapid evidence assessments integrate literature evidence to produce a narrative summary. Systematic reviews often take months and therefore, rapid evidence assessments can be used if less rigour can be tolerated. These methods often focus on quantitative data, but other frameworks can be used to systematically integrate qualitative information. Guidance on rapid evidence assessments[50] and databases of systematic reviews are available[51,52].

A set of questions is provided below in Figure 3.2 to help consider which of these evidence synthesis methods are likely to be the most appropriate. Table 3.5 provides an overview and pros and cons of the most common evidence synthesis methods. More detailed information can be found in Annex A.

50 Civil Service (2016) *Rapid Evidence Assessment: Resources and Guidance*.
Available at: https://www.gov.uk/government/publications/the-production-of-quick-scoping-reviews-and-rapid-evidence-assessments (Accessed: April 2026).
51 Cochrane Library (2019) *Cochrane Database of Systematic Reviews*.
Available at: https://www.cochranelibrary.com/ (Accessed: April 2026).
52 Campbell Collaboration (n.d.) *Better Evidence for a Better World*.
Available at: https://www.campbellcollaboration.org/evidence/ (Accessed: April 2026).

Figure 3.2: Selecting evidence synthesis methods

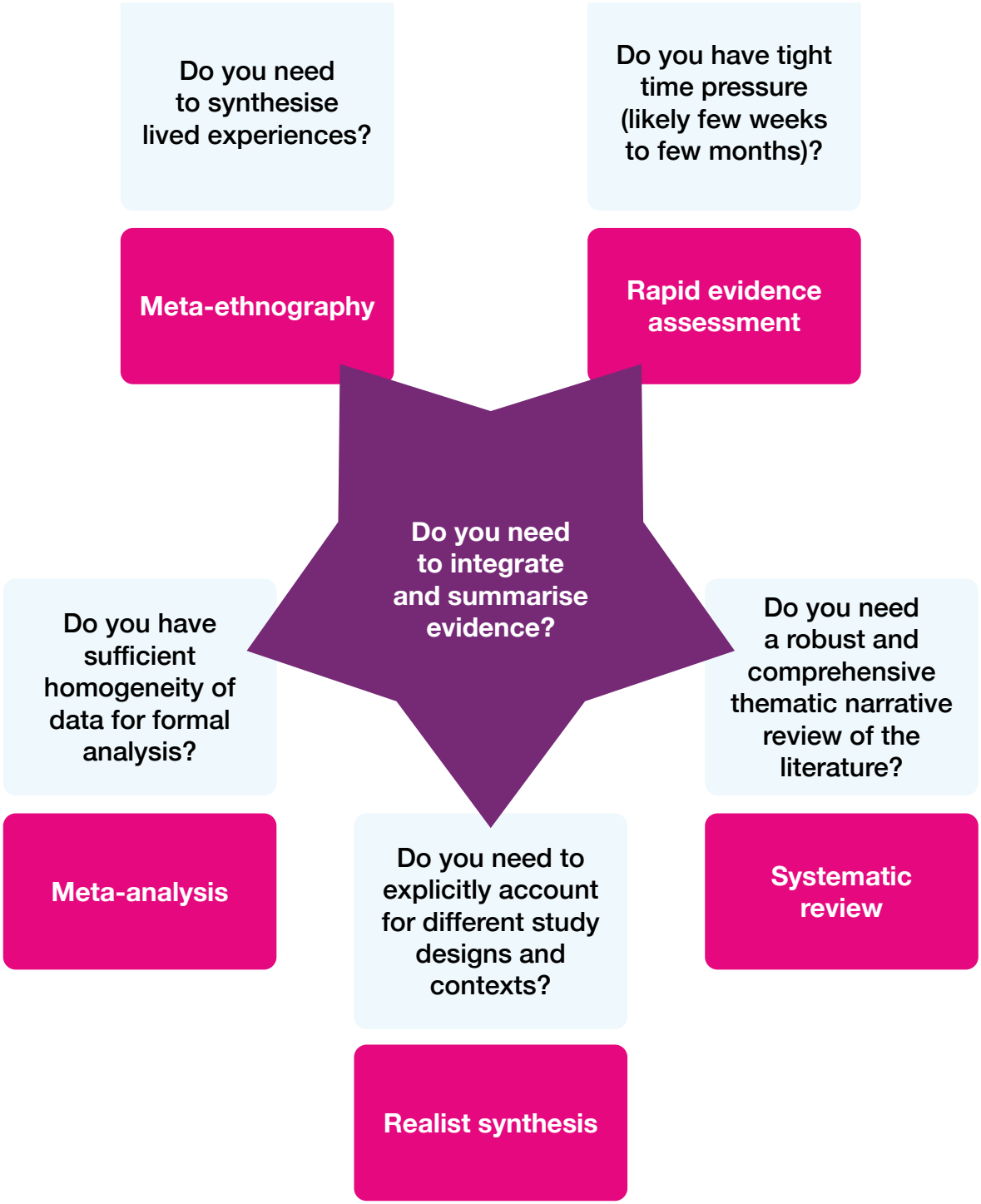


Table 3.6: Quick guide to common evidence synthesis methods

Evaluation method	Description	Pros and cons
Rapid evidence assessment	Rapid evidence assessments are systematic but relatively quick literature reviews. They may use interviews with experts to facilitate targeted searches of the literature. The focus tends to be relatively narrow.	Useful for relatively quick (approx. 3 months) scoping out of existing evidence but require a number of good quality studies to draw from. They are less effective where research questions do not easily map on to the existing body of evidence. Transparency is key. They are subject to greater bias than a systematic review.
Systematic reviews	A systematic method of identifying, assessing, extracting and integrating evidence from multiple studies. The focus should provide a methodical approach to reviewing all data to answer a clearly defined research question. The aim is to reduce bias and provide a comprehensive overview of studies through use of five key steps.	Provides a comprehensive assessment of available evidence. Has a rigorous approach to assessing and referencing studies. Can be resource intensive and take a long time (six months plus). They need a substantial body of evidence to review, so may be less effective when used in fields with limited evidence.

Evaluation method	Description	Pros and cons
Meta-analysis	Meta-analyses strengthen the evidence of the impact (or lack of impact) of an individual intervention. They statistically integrate quantitative results from primary studies to provide reproducible summary estimates, such as the average effect size of an intervention. They can be based on a pooling of the individual observations or the average effect sizes from the original studies. Interventions analysed may differ in content or delivery and the studies of them may vary in their methodology but these may be addressed in analysis.	Can improve estimates of the size of effect of an intervention or resolve uncertainty between studies of similar interventions. The quality of the meta-analysis is dependent on the quality of the individual experimental studies. Quantitative source data must be equivalent and sufficiently consistent. Validity of the method depends on the inclusion criteria. Meta-analyses of quasi- experimental studies may not account for bias and confounding factors.
Meta-ethnography	Meta-ethnography brings together multiple individual lived experiences. Evaluators select, analyse and interpret narrative accounts within studies to identify new concepts and insights across studies.	Allow evaluators to either evolve overall concepts, explain conflicting differences in stories, or build a picture of the study subject from a number of different narrative accounts. Quality depends on quality of original studies and validity depends on inclusion criteria.
Realist synthesis	This integrates systematic review methods with realist evaluation theory to explain outcomes from an intervention through literature review. It is often used to assess complex policies based in complex environments as it is a structured method to understand the underlying context-mechanism-outcome relationships in an intervention.	Accounts for differences between study contexts and study designs. The broad concepts to the method are specific to the intervention and context and therefore cannot easily be reproduced or standardised. Requires researcher knowledge of the intervention context and understanding of implementation.

HM Treasury

Chapter Four

Data collection, data access and data linking

4

Summary

The collection of data required for an evaluation should be planned alongside the development of the intervention; where this does not occur, an evaluation may be impossible, severely limited or unnecessarily expensive.

In planning data collection, the following should be considered:

- The evaluation questions to be answered.
- Who can provide relevant data.
- Data access constraints.

Existing data from administrative and monitoring systems or large-scale (long-term) surveys are important sources for evaluation. They can be particularly valuable for providing longer-term trend information that pre-dates the evaluation.

Typically, new data collection and research is also required.

Maximising data quality can be achieved through minimising bias, testing data collection processes in advance, and building in pre-planned data quality checks.

Data must be handled appropriately; data access protocols must ensure compliance with data protection regulations.

Linking different data can create richer datasets, improve the quality of data and avoid duplication of data collection. However, doing this presents data management, ethical and analytical challenges.

4.1 Introduction

Data collection is an essential component of any evaluation and needs to be planned in advance. Planning data collection alongside an intervention's development ensures that data collection and data access are built into the policy's design and related legislation.

It is important to note that baseline data – data collected before the intervention – will need to be collected early, and that comparison data – data from groups unaffected by the intervention – will need to be negotiated. Without appropriate data collection or data access planning an evaluation may be impossible, severely limited, or unnecessarily expensive. If data collection is poorly designed, inaccurate data may be collected and false inferences drawn from the evaluation.

The Theory of Change (see section 2.2.1), can be used to identify data needs and gaps. A combination of the collection of existing and new data may often provide the most complete understanding and tracking of the Theory of Change and effectiveness of the intervention.

Collecting data using several methods, such as monitoring data and bespoke surveys, builds confidence in the robustness of the evaluation and its findings – this is known as **triangulation**.

In many cases, data relates to individual people but can relate to other units, such as schools, businesses and geographical areas. Similar considerations apply in all cases.

4.2 Deciding what data is required

The choices are not straightforward in determining the data required to answer the evaluation questions – input should be sought from those with relevant expertise.

Table 4.1 below outlines the key questions to consider on data collection when planning an evaluation.

Table 4.1: Deciding what data is required

Key question	Considerations
What type of data will be required to answer each evaluation question?	What is needed to answer the evaluation questions identified during the scoping stage. Types of data include: • Numerical data. • Documentary data (data or information that has already been collected). • Observational data. • Descriptions of people's experiences, opinions and views. • A combination/triangulation of the above evidence.
Who or what can provide this data?	• Intervention participants. • Service providers. • Stakeholders. • Databases. • Existing surveys. • Bespoke surveys.

Key question	Considerations
Are there any issues with accessing or collecting the data?	• Data access issues (such as legal issues, internal procedures, identifying the target group(s), collecting comparator data). • Data sensitivities/ethical issues (such as researching sensitive populations, data access issues). • The availability of necessary sampling frames. • The potential for 'data burden' on respondents: is it proportionate to ask people to provide the data? • Who is responsible for data collection? • Is the data collection task proportionate to its value? • Are other methods possible? • Could the act of collecting the data from individuals influence their behaviour? • Could the data collection approach affect accuracy or create a risk of bias?
What section of the population of interest should data be collected from?	There are a number of ways to collect data from the population of interest. It is important to consider what is proportionate to answer the evaluation questions. Options include: • A census of all available data/populations of interest. • A representative sample of the data available/population. • A subset of the population of interest, purposively selected to cover a range of contexts but not statistically representative.
How will the proposed data analysis method influence the data required?	The type of analysis that will be applied to the data will typically dictate the nature and quantity of the data required. For example: • The required degree of precision. • The sub-populations of interest (and therefore, the sample size required). • The nature of the method to be used and therefore, the data requirements. • Whether baseline or control group data is necessary.

4.3 Sources of data

There are many different sources to consider when seeking to collect data for an evaluation. These include:

1. **Existing administrative and monitoring data**: data gathered either for the operation of an intervention, or for other functions.
2. **Existing large-scale survey data:** long-term, large-scale survey data, often managed by central governments, the Office for National Statistics or research funding bodies, such as UKRI (such as data from the Labour Force Survey or Crime Survey for England and Wales).
3. **New sources of data designed specifically for the evaluation:** such data are typically obtained through methods such as surveys, qualitative methods (interviews, observation, focus groups), web-scraping.
4. **Social media data:** potentially very rich in terms of gauging an unprompted reaction to an intervention, but researchers should be aware of potential biases in the data.

Where possible, existing administrative or other data should be used as this will be cost-effective and often covers the whole population of interest. However, new data are often required to answer specific questions, in particular those not essential for performance management of a programme.

A wide range of both quantitative and qualitative data collection methods can be used. Each data collection method will now be described in turn covering its purpose, application and considerations in an evaluation.

4.3.1. Administrative/monitoring data

Administrative data

Administrative data is collected, stored and used primarily for administrative (non-research) purposes. Administrative data is typically collected on registrations, transactions and record-keeping, usually during the delivery of a service (for example: the National Pupil Database held by DfE; the Police National Computer held by MoJ; and records from service providers on their customers and the services provided). Administrative datasets are commonly used to form the basis of Official Statistics' publications.

The availability of administrative data or data from an existing long-term survey should always be considered at the design stage of an evaluation. Both types of data can serve as important sources of background or explanatory data that pre- date the intervention giving excellent trend information. Where possible, the use of existing data reduces both the financial costs of the evaluation and the burden on respondents.

However, care needs to be taken when using this data for evaluation purposes as the data might not focus on the specific questions of interest nor be collected regularly enough to provide useful evidence against the evaluation questions.

In some cases, administrative data measure concepts that are related but not identical to the measures of interest. For example, while administrative data can tell us with great precision and accuracy who is in receipt of benefits, measuring the most widely accepted definition of 'unemployment' requires survey data. Although these two concepts are closely related, they are not the same.

AI and machine learning can accelerate the analysis of large administrative datasets by identifying complex patterns or predictive indicators that may not be immediately apparent. While this can help generate new lines of inquiry for an evaluation, any such findings must be rigorously tested rather than taken as causal evidence, in line with the principles outlined in section 5.6.

Monitoring data

Monitoring data or performance management data are collected throughout an intervention to provide answers to policy, research and performance questions.

Monitoring data typically cover all aspects of an intervention's operation, for example, information about those accessing a service, as well as the project's inputs, processes, outputs and outcomes (see Table 4.2 below on how monitoring data can be used in evaluation). Monitoring data is frequently administrative and quantitative, and often not generated primarily for evaluation purposes. It is generally used to help track progress of an intervention's delivery; or to identify where an intervention is not being implemented as expected and what further action is required to ensure it can achieve its objectives. Monitoring data can, if suitably defined and collected, be used to measure benefits as part of a formal benefits management process. Guidance on effective benefits management is available from the Infrastructure and Projects Authority[53] and other sources.[54]

53 Infrastructure and Projects Authority (2017) *Guide for Effective Benefits Management in Major Projects*. London: Crown Copyright. Available at: https://www.gov.uk/government/publications/guide-for-effective-benefits-management-in-major-projects (Accessed: April 2026).
54 Jenner, S. (2014) *Managing Benefits*. 2nd edn. London: The Stationery Office.

Table 4.2: How monitoring data can be used in an evaluation

Monitoring data	Example	How this can be used to evaluate
People accessing a service	Numbers and characteristics.	Used to assess whether an intervention is reaching its target population and to what extent, features of that population and whether non-target groups are also being affected touched.
Inputs	Funding, resources, staff numbers.	Assesses whether the inputs required met expectations. Used to inform cost-benefit analysis and determine whether assumptions about policy implementation, such as cost and time, were correct.
Processes/ activities	Referrals, waiting times.	Used to determine whether the policy is being implemented as planned or whether there are any unintended consequences.
Outputs	Numbers going through a programme/ application processed.	Used to inform an assessment of whether the programme has delivered the target outputs to the anticipated quality.
Outcomes	Employment rates, wages.	Used to measure the benefits of delivering the outputs.

Figure 4.1: Developing a monitoring system

What data needs to be gathered to give reliable and consistent measurement at each stage of the Theory of Change (and if necessary evaluate impact)?

1. Identify key indicators that can be clearly defined and used to monitor progress against the Theory of Change. Targets and target dates can be set but are not essential.
2. Identify the data required to measure inputs, outputs, outcomes and impacts.
3. Identify data needed to create a comparison group (if needed for an impact evaluation).
4. Identify whether this data is already available.

What new data should be collected to fill gaps in measurement?

The following data should be considered:
1. Stakeholder perceptions/attitudes towards the intervention/behavioural change
2. Financial data relating to the intervention's expenditure
3. Process data to assess whether the intervention has been implemented as intended
4. Data to track the outcomes and impact of an intervention

Who will have responsibility for gathering the data?

1. Identify the most appropriate individuals to gather the data, e.g. programme/project delivery team, an existing performance monitoring team, evaluators, etc.
2. What resources are required? Do those responsible have the time and skills?

When will the data be gathered?

1. How often should administrative/monitoring data be gathered? (e.g. monthly, quarterly, annually). Can this be aligned with the auditing/reviewing process of the funding body?
2. What is the timetable for the collection of new data and is this aligned with the reporting schedule for the evaluation?

How will the data be gathered and stored?

1. What format should the system use? Can it be aligned with existing monitoring systems?
2. Data protection protocols are needed to meet security and data sharing requirements.
3. Likewise, ethical considerations need to be taken into account (e.g. informed consent).

How will the data be gathered and stored?

1. What format should the system use? Can it be aligned with existing monitoring systems?
2. Data protection protocols are needed to meet security and data sharing requirements.
3. Likewise, ethical considerations need to be taken into account (e.g. informed consent).
4. Where will the data be stored?

How will the data be verified to ensure accuracy and consistency with requirements?

1. Who are the most appropriate individuals to verify the data, e.g. analyst, programme/project lead at the funding body, independent evaluators, etc.?
2. What resources are required to undertake the task?

Design and implement the monitoring system

Monitoring data can play a key part in the evaluation of an intervention by providing useful data throughout the life of a policy, and by providing the data with which to conduct some of the analytical methods discussed in Chapter 3. The quality of monitoring data is often a function of the value that those responsible for collecting the data can see from the task. If it is seen purely as an administrative burden, the incentive to ensure data quality is typically low – if it is used directly by the service providers, the incentive to ensure quality is strong. Data quality is considered further in section 4.4.

Key questions and considerations need to be considered to inform the design of an effective monitoring system. An effective monitoring system needs to be designed before activity starts on the ground, ensuring that necessary data are collected (such as baseline data), and should be linked to the Theory of Change for the intervention. Figure 4.1 above illustrates these key questions and considerations.

4.3.2. Survey data

Large-scale survey data is often collected by government departments and research agencies for statistical and research purposes. Examples of these include: the Childcare and Early Years Providers Survey by DfE; the HMRC Customer Survey,[55] and Understanding Society by the Economic Social Research Council (ESRC).[56]

Large-scale survey data typically has large representative samples, which provide robust estimates. Using these surveys can be cheaper than collecting new data. However, there are potential downsides:

- The purpose for the data collection will be different to the purpose of the evaluation.
- They are unlikely to give the detail required; that is, answers to specific questions or information about the populations of interest (for example, the total survey sample may be large, but the number of people in the sample targeted by the intervention may be small and it may not be possible to accurately identify them).
- The timing of the surveys will be fixed and might not suit the evaluation's purposes.

They are, however, often excellent sources of background or exploratory data, for example, providing the prevalence of specific activities/behaviours/attitudes in the general population, which can be useful in framing the problem or explaining the context.

55 GOV.UK (2019) *GOV.UK Official Website*. Available at: https://www.gov.uk (Accessed: April 2026).
56 Understanding Society (2019) *Understanding Society Official Website*. Available at: https://www.understandingsociety.ac.uk/ (Accessed: April 2026).

New quantitative data collection usually involves surveying a sample of the population to make estimates about the broader population. It may be appropriate to conduct a census of the entire population; or in some instances, responding may be made mandatory as part of the conditions of taking part in an intervention. In these instances, there will be less likelihood of non-response bias.

Sampling

Sampling methods fall under two broad categories: probability sampling and non-probability sampling.

Most surveys used in evaluations will be based on probability sampling methods, which involve selecting respondents at random from a sampling frame (for example, a list of all respondents in the population of interest). The main methods for deriving probability samples are:

1. **Simple random sampling:** Involves using a random method, typically computer-generated, to select individuals, with all individuals having an equal probability of selection.
2. **Stratified sampling:** Dividing a population into groups then selecting a probability sample from each group, for example, based on geographical location. The probability of selection can vary between groups to ensure that there is a sufficiently large sample from each.
3. **Cluster sampling**: Clusters are selected using a probability method and only individuals within clusters are selected. This can help to reduce fieldwork costs, primarily in face-to-face surveys where the clusters are typically geographic areas but does increase the uncertainty of estimates derived from the data.
4. **Multi-stage sampling:** Involves using a combination of sampling methods.

Providing there is not a problem with non-response (more on this at the end of this section), this type of sampling will ensure that estimates derived from the survey are representative of the population as a whole (or can be weighted to make it so), and that the sampling error of those estimates can be calculated.

Non-probability approaches can be convenient and less costly but should be used with great caution as it is not possible to infer with any confidence anything about the population as a whole from the survey results.

Quota sampling is a widely used non-probability approach, which involves setting 'quotas' for different types of respondents. This allows for greater control over the sample composition, such as ensuring adequate representation of different groups, or over-sampling of minority groups who are not necessarily identified in the sampling frame. Well-designed quota samples are a valid alternative to probability sampling. The strongest quota sample designs start with a sample that is drawn at random from a representative sampling frame and then set quotas for the number of interviews to be achieved with particular groups. It should have fieldwork procedures which mitigate the risk of only including those respondents who are easiest to get a response from. Weaker designs can involve 'convenience samples' of individuals who are easy to contact, or 'snowball samples'. Further government guidance on quota sampling is available.[57]

All survey designs will need to consider how to maximise response rates and minimise non-response bias, such as through fieldwork protocols that specify the number of attempts that should be made to contact people.

57 Government Statistical Service (2018) *Quota Sampling Guidance*.
Available at: https://gss.civilservice.gov.uk/wp-content/uploads/2018/03/Quota-sampling-guidance-4.pdf (Accessed: April 2026).

If levels of non-response are not evenly distributed in a probability sample, other than as related to characteristics observed in the sampling frame, this may lead to biased results, and the confidence in inferences about the wider population will be reduced.

Survey questions

Survey questions can be used to collect different types of information; these are covered in Table 4.3 below.

Table 4.3: Types of survey questions

Types of questions	Type of information collected
Factual questions	Surveys often offer the only practical and affordable way of collecting such information, and in some cases, there is no other source or way of measuring the attribute of interest. This can include both objective and subjective measures.
Knowledge questions	Assess what respondents know about a particular topic and their awareness of the intervention being evaluated.
Attitudinal questions	Measures respondents' opinions, beliefs, values and feelings which cannot be verified by observation or external data sources.
Behavioural questions	Measures what people do, or intend to do, and how that has changed as a consequence of the intervention. A risk here is respondents giving socially acceptable answers (although, good survey design and experienced interviewers can minimise this). Triangulating with observed behaviour can be useful.
Preference questions	Respondents provide preferences for different possible options and outcomes, including trade-offs between competing policy objectives. These can be used to elicit monetary values for different outcomes, including those not readily possessing market prices (such as changes in air quality, health status) for use in cost-benefit analyses.

When designing surveys there are four rules that are useful to consider:

1. Can the respondents understand the question – and do they understand it in the same way that you do?
2. Are respondents able to answer the question?
3. Are they willing to answer the question?
4. Will the question produce a reliable and useful response?

Most data collection tools and their associated materials (such as show cards) will require development, possibly involving cognitive testing – researching how people understand and respond to interview questions – or pilots, to ensure the questions are fit for purpose. Where possible, using the same questions used in other surveys can reduce the need for testing, allow for comparison between surveys and help build the wider evidence base.

Standard formats for questions include:

1. **Harmonised questions:** The Office for National Statistics (ONS) 'Harmonisation' programme has been working towards harmonised questions[58] in areas such as age, gender and ethnic origin.
2. **Validated questionnaires:** Using validated questionnaires – such as the GHQ-12 set of questions measuring mental wellbeing and the EG-5D questions measuring health status – enables comparisons with other studies and ensures the results of the evaluation can be correctly interpreted.
3. **Question Bank:** The UK Data Archive[59] maintains a Question Bank, which catalogues questions that have been used in previous surveys.
4. **Maintaining consistency between surveys:** Repeating surveys at the same time of year, in the same geographical areas, or using the same sampling frame ensures the results between surveys are comparable and not subject to seasonal variation or other factors not related to the intervention.

There should be a presumption from the outset that all survey data should be archived, if possible, commonly through the UK Data Service.[60] This will typically require additional work to anonymise the dataset, which should be built into any contract for commissioned survey work.

If there are plans to retain personal information to enable future follow-up, informed consent and data management processes need to be built into the survey to allow this. Contracts should also consider copyright of intellectual property, including the questionnaire and datasets.[61] Legal requirements that need to be met for the processing of personal data under General Data Protection Regulations (GDPR) are discussed in section 4.5.[62]

58 Office for National Statistics (ONS) (2019) *Harmonisation with the GSS*.
Available at: https://www.ons.gov.uk/methodology/classificationsandstandards/harmonisationwithinthegss (Accessed: April 2026).
59 UK Data Service (2019) *Question Banks*. Available at: https://www.ukdataservice.ac.uk/get-data/other-providers/question-banks.aspx (Accessed: April 2026).
60 UK Data Service (2019) *UK Data Service Official Website*. Available at: https://www.ukdataservice.ac.uk/ (Accessed: April 2026).
61 UK Statistics Authority (2019) *Code of Practice for Statistics*. Available at: https://code.statisticsauthority.gov.uk/ (Accessed: April 2026).
62 GOV.UK (2018) *Guide to the General Data Protection Regulation*.
Available at: https://www.gov.uk/government/publications/general-data-protection-regulation-policy. (Accessed: April 2026).

4.3.3. Qualitative data

Qualitative data collection methods provide an in-depth understanding of behaviours, perceptions and underlying reasons for social phenomena. While quantitative methods are usually used to measure the 'what', qualitative methods are most often used to explore the 'how' and 'why'.

Some common qualitative data collection methods.

1. **In-depth interviews:** These are used for collecting data on individuals' personal histories, perspectives and experiences, particularly when sensitive topics are being explored, or the issue being discussed is not well understood. Questions tend to be 'open' allowing for detailed responses. In-depth interviews are typically face to face or by telephone.
2. **Focus groups:** Focus group participants are encouraged to discuss and debate openly to understand views and experiences, allowing for a range of views to be explored. These can be used effectively for action planning and developing or improving products and services.
3. **Case studies:** These are in-depth, possibly longer-term, investigations of a single issue or a small number of people, events, contexts, areas, organisations or policies.
4. **Observation:** A process of watching research subjects (with their agreement and knowledge) to observe their behaviour, without questioning them. Can be used effectively when piloting new procedures or processes to gauge respondents' actions and behaviours.
5. **Ethnography:** This includes observation and participation. Ethnography seeks to understand people and how they live in their cultural and physical environment. Ethnographic interviews will differ from more traditional in-depth interviews as the researcher would usually have shared time and built relationships and trust with the interviewees. Film ethnography is one example of how this approach is being used by Policy Lab and the government.[63]

Selecting respondents

For qualitative data collection, the selection of respondents aims to ensure the data captures the richness of views and opinions in the population of interest – or by illustrating specific stories or contexts through very in-depth exploration. It is important to consider which viewpoints should be included in the evaluation as this will impact on the sample size.

There are no agreed rules of thumb for how many respondents should be selected. The objective is not to be able to produce findings that are statistically representative of the population, but to understand the range of views on a particular topic, and to provide deeper insights into social phenomena. Reaching saturation point – the point at which the addition of new data doesn't add anything new to the findings – is a good indication that the breadth of views and opinions has been captured and additional interviews are unlikely to add value.

There are additional sources of information that can assist in selecting a qualitative sample, such as that published by the National Centre for Research Methods (NCRM).[64]

63 Andrews, B. (2018) 'Positive Engagement – Our User-Centred Approach' [blog], *Policy Lab*. Available at: https://openpolicy.blog.gov.uk/2018/07/25/positive-engagement-user-centred-approach/ (Accessed: April 2026).
64 Baker, S. (2012) *How Many Qualitative Interviews is Enough? Expert Voices and Early Career Reflections on Sampling and Cases in Qualitative Research*. Available at: http://eprints.ncrm.ac.uk/2273/4/how_many_interviews.pdf (Accessed: April 2026).

4.3.4. Social media data

The use of social media as a source of data for evaluation has been increasing in recent years, both for qualitative and quantitative analysis. How social media is used as a data source is determined by the types of behaviour being researched and the platforms that are used as source material. The development of automated tools (web scraping) allows for large volumes of data to be collected, cleaned, stored and analysed quickly. The use of social media data presents special challenges. Samples, while very large, are self-selected, and it is not always easy to assess the reliability likely accuracy of the data.

Further guidance on the use of social media data is available from GSR.[65]

65 Social Media Research Group (2016) *Using Social Media for Social Research: An Introduction*. Government Social Research. Available at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/524750/GSR_Social_Media_Research_Guidance_-_Using_social_media_for_social_research.pdf (Accessed: April 2026).

4.4 Data quality

Data quality is of prime importance to producing robust findings that can be clearly and appropriately interpreted by decision-makers. The quality of existing data should be assessed at the outset – poor quality or partial data will affect the scope and scale of the data's contribution to an evaluation. Transparency is key to this and is discussed further in section 6.6. Quality checks should be built into the evaluation design, and this is discussed in section 5.8.

Data quality will depend on the type of question(s) asked and the type of data collection tool used.

Data about potentially sensitive issues, such as sexual orientation or disability, where different people can interpret the question in different ways, are often more reliable when collected as part of a research exercise than through monitoring systems.

On the other hand, administrative or monitoring data will be of higher quality when examining issues such as exact dates of joining or leaving a programme. Such data is more likely to be recorded accurately on monitoring systems rather than be recalled by a participant in an interview. In some cases, usually when used directly as part of a payment system, data are subject to formal auditing and are therefore, particularly reliable.

The Aqua Book[66] and the Green Book[67] provide further guidance on producing quality analysis for government. Additional guidance on quality in qualitative evaluation is outlined in Annex B.

66 HM Treasury (2025) *The Aqua Book: Guidance on Producing Quality Analysis for Government*. London: Crown Copyright. Available at: https://www.gov.uk/guidance/the-aqua-book (Accessed: April 2026).
67 HM Treasury (2026) *The Green Book*. London: Crown Copyright. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/685903/The_Green_Book.pdf (Accessed: April 2026).

4.5 Data handling

All data must be appropriately collected, handled and accessed. Everyone handling data must be suitably trained to understand how to do this. There are special considerations under GDPR[68] where personal data[69] is concerned.

All data used in an evaluation, regardless of its source, must be collected, transferred, stored, processed and deleted in accordance with GDPR and any specific departmental security processes. This holds for external organisations, such as research contractors, collecting data on a department's behalf. Data access protocols should cover matters such as how to authorise access to data in different groups: remote access, Wi-Fi access, data handling and security training, and masking/encrypting data.

Data can be stored in many different formats, such as in a database, dataset, spreadsheet or data warehouse. Ensuring appropriate data access – those able to read, write, store and change data – is critical and should be restricted to those who have legal entitlement and a work requirement.

Data handling risks are serious and include criminal data breach, individuals or groups suffering severe harm or embarrassment, unjustified intrusion, loss of privacy, data loss or compromise, legal challenges, and reputational damage at department, Civil Service and wider government levels.

Data should be anonymised at the first opportunity – that is, direct identifiers such as names and addresses should be removed from the analytical file. If it is necessary to retain these, at least temporarily, they should be stored separately with strictly controlled access. Even so, in most cases the data is likely to remain personal data and should be carefully protected. Implications for the informed consent process are discussed in section 5.9.

At the end of a project, if possible, data should be fully anonymised: a process that is likely to include, for example, replacing precise date of birth with month and year of birth, or exact postcode with a broader geographic coding. If this is not possible, and it is deemed necessary or desirable to retain personal data beyond the life of the project, appropriate safeguards must be taken.

68 Information Commissioner's Office (2018) *Guide to the General Data Protection Regulation*. Available at: https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/ (Accessed: April 2026).
69 GDPR defines 'personal data' as any information relating to an identifiable person who can be directly or indirectly identified in particular by reference to an identifier.

4.6 Data linking

Linking different data can create much richer datasets, enhance analysis, improve the quality of data and avoid duplication. It can allow evaluators to answer complex questions in a cost-effective way. However, the process of linking datasets is not straightforward and creates data management, ethical and analytical challenges.

Data linkage combines different sets of data that have been collected for different purposes. It can be the combination of two or several datasets and link survey data with operational or administrative data. Data linkage does not necessarily involve public administration data and can also be performed with data collected by private sector or academic organisations.

In the current 'big data' and 'open policymaking data' context, the potentials of linking data have expanded. Recent developments, including the introduction of the Digital Economy Act and the establishment of Administrative Data Research UK, should make data linking more feasible than it has been in the past. These are discussed below.

4.6.1. How data linkage works

For datasets to be linked, it is necessary that they include identifiers that will allow records to be paired. Some datasets will include unique identifiers, such as National Insurance numbers or NHS numbers. Firms can also be linked by VAT registration by using data from Companies House. In these cases, linkage can be straightforward.

In the absence of unique identifiers, linkage can happen using information, such as name, date of birth and postcode. As these can be imperfect and use different formats, 'fuzzy matching' is typically used to maximise the number of links made. AI- and machine learning-driven techniques may be able to improve the process of 'fuzzy matching' when unique identifiers are absent, potentially improving the accuracy and speed of linking large and complex datasets. However, inevitably, there will be a proportion of false positives (records that correspond to the same person fail to link) and false negatives (unrelated records linked by mistake). Careful thought needs to be given to the matching protocols to achieve an appropriate balance of these, and to the implications for the analysis of the linked data.

It is important to ensure datasets are capturing the same cases in an accurate way and are consistently recording identifiers to ensure data quality. For example, research on data linkage of medical records found that a number of relevant patient or population factors may be associated with incomplete data linkage resulting in systematic bias in reported clinical outcomes.[70]

70 Bohensky, M.A., Jolley, D., Sundararajan, V. et al. (2010) 'Data Linkage: A Powerful Research Tool with Potential Problems', *BMC Health Services Research*, 10, p. 346. doi: 10.1186/1472-6963-10-346.

4.6.2. The legal framework in the UK

Data linkage raises issues around access to personal data, data protection, consent and ethics, especially when the data to be linked is collected and held by government.

To facilitate access to linked government data by researchers while ensuring individuals' privacy is protected, many countries have created data linkage research centres and initiatives. The model is of a 'trusted third party' as these centres allow researchers to access de-identified linked data in a secure environment. In the UK, the Digital Economy Act (2017) introduced powers to authorise data processors to receive, link and enable research access to administrative data. These powers, including the accreditation of data processors, researchers and research projects are managed by the UK Statistics Authority, working with data-owning departments.

The ESRC has established Admin Data Research UK,[71] a partnership between a number of data processors (including ONS), to facilitate and promote access to government-owned data. The partnership has a particular focus on supporting research that is aligned with strategic priorities across government while aiming to minimise the burden on departments.

71 ADRU (2019) *ADRU Official Website*. Available at: https://www.adruk.org (Accessed: April 2026).

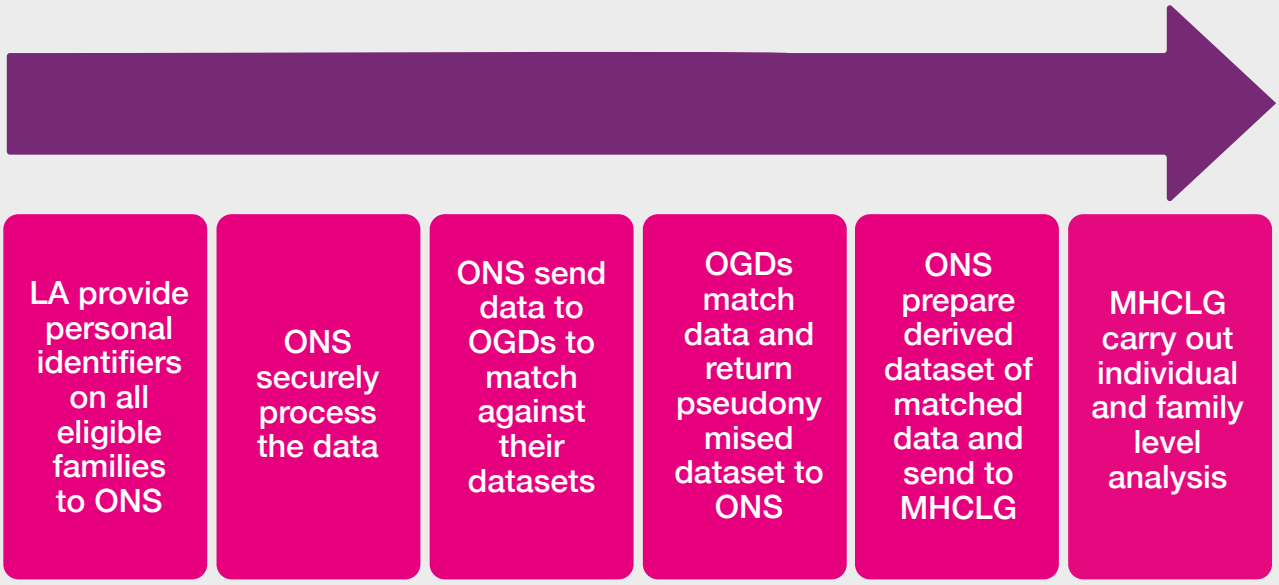
Figure 4.2: Data linking case study

Troubled Families Programme (Ministry of Housing Communities and Local Government, MHCLG)

The evaluation of the second Troubled Families Programme (which ran from 2015–2020) uses individual-level data from nationally held administrative datasets linked together to measure the impact of the programme (National Impact Study, NIS). The diagram below illustrates the data linking process.

NIS Data linking

Personal identifiers (names, D.O.B, addresses, gender) were collected from local authorities (LAs) and used to link to administrative data held by other government departments (OGDs).



Local authorities were asked to supply the Office for National Statistics (ONS), acting as a data processor on behalf of MHCLG, with personal information (names, addresses, dates of birth and gender), including the local authority ID number, of all individuals eligible for the programme. This included all families participating in the programme, as well as a comparison group in each area.

The ONS replaced the local authority ID number with an ONS ID number, cleaned the data and prepared a dataset of ONS ID numbers and personal information for each government department providing data for the evaluation. Each department used the personal information provided to identify individuals within their administrative datasets and create a pseudonymised dataset (a dataset of matched data that only included the ONS ID, for example, no personal information). The pseudonymised data was returned to ONS. The ONS used the raw matched data to create a dataset of derived variables ready for analysis.

This was sent to MHCLG analysts and used to carry out individual-level and family-level analysis on the anomymised data, as well as estimate the impact of the programme on a range of outcomes. All published findings from the research were based on aggregate statistics and presented in a way that did not allow individuals or families to be identified.

The data linking is governed by Data Sharing Agreements with each local authority and memorandums of understanding (MOUs) agreed with each of the government departments (Ministry of Justice for Police National Computer data, Department for Education for National Pupil Database, and Department for Work and Pensions/Her Majesty's Revenue and Customs for Work and Pensions Longitudinal Study data). A Data Privacy Impact Assessment (DPIA) was prepared. Lawyers and data security experts were involved in drawing up these documents and advice on how to meet the requirements of the data protection legislation was provided by the Information Commissioner's Office (ICO).

To get MOUs in place took two years of negotiation with key partners, including local authorities and government departments. The MOUs allow data retention for two and a half years after the programme has ended, to facilitate a longitudinal study to look at whether outcomes are sustained. These negotiations included consideration of the legal gateways for sharing data, ethical concerns (privacy notices were issued rather than relying on informed consent), and data security measures for each part of the project data flows. Data security measures had to meet pan-government security standards.

HM Treasury

Chapter Five

Managing an evaluation

5

Summary

It is crucial to set up effective governance structures at the beginning of an evaluation to ensure the evaluation is fit for purpose, can adapt to changing situations, has quality assurance and stakeholder buy-in, and can influence decision-making.

Resource requirements, both financial and human, should not be under-estimated. Managing an evaluation requires both specialist evaluation skills and general project-management skills.

Different routes to conducting an evaluation are available: external, in-house or a combination of the two.

Regardless of the route chosen, the specification is critical to ensuring the quality of any evaluation. Specifications may be tightly or loosely defined; there are advantages and disadvantages to both approaches.

Flexibility in the evaluation is often required and must be built into the specification and management processes.

Ethical issues need to be considered throughout.

5.1 Introduction

This chapter provides guidance on different aspects of managing an evaluation, including governance, linking to the intervention design, specifying and commissioning, managing and assuring the quality of the evaluation alongside wider issues, such as ethics and resources.

5.2 Establishing the evaluation

The best environment in which to plan an evaluation is one in which there is an expectation for all interventions and policies to be evaluated. When this is not the case, ensuring a commitment to an evaluation should be achieved as early as possible in the process. As set out in Chapter 2, evaluation should be explicitly considered at all stages of business planning, and resources built into the business case to cover both internal expertise and external spend.

Evaluation can require substantial resources; both financial and from analysts and policymakers. Chapter 7 describes in more detail the capabilities required. The main resources and the considerations that should be made are described below in Table 5.1.

Table 5.1: Resources required in an evaluation

Resource	Considerations
Financial	There is no standard proportion of an intervention's value that should be dedicated to evaluation, although the scale of the evaluation is generally proportionate to the scale and ambition of the intervention. Pilots, trials and other new policies will spend more (proportionately) on evaluation than routine, well-established policies. Collecting new data often accounts for a large proportion of the cost of an evaluation. Building data requirements into routine monitoring activities can, therefore, substantially reduce costs.
Management	Within government departments, evaluations require a project manager to be responsible for planning, commissioning, day-to-day management, and issues resolution. This is most efficient when the project manager has appropriate analytical skills to quality assure and resolve technical issues. Government project managers should also ensure that those designing or conducting the evaluation remain independent and unbiased and do not allow unacknowledged assumptions about the intervention to interfere with their work. It is useful for the project manager to work closely with the policy or delivery team to ensure the evaluation is of maximum value.
Analytical (including external researchers conducting the evaluation)	Evaluation is a multi-disciplinary activity and can typically require a range of specialists to advise on its design and outputs, including evaluation specialists, social researchers, statisticians, operational researchers and economists. The value of an evaluation specialist should not be underestimated, particularly for large or complex evaluations. The design and execution of an evaluation is often a complex and technical task. Multi-disciplinary support is also useful to steer and quality assure the evaluation. External peer reviews are typically commissioned, and therefore, need to be resourced and given a lead-in time. Often, external researchers who specialise in evaluation will undertake the evaluation. It is important that they fully understand the requirements of the evaluation, but in some cases, it can be useful for them to maintain a degree of independence from commissioners to produce confidence in the evaluation findings.

Resource	Considerations
Policy	Engaging with the evaluation is essential for the policy or delivery team responsible for the intervention, and they should allocate resource accordingly. This will include day-to-day engagement to ensure the evaluation is focused and the findings are useful, and through discussions at policy/programme boards where appropriate (See section 5.3).
Project delivery/ Project Management Office (PMO)	Benefits management is a project delivery discipline focused on defining and measuring the desired changes or benefits of an intervention. Benefits management and evaluation approaches may interact and inform one another. At the start of the project, it can be useful to ensure consistency with the evaluation – particularly where similar benefits, outputs or outcomes are being measured in both. You should also consider whether joined-up reporting would be beneficial.
Delivery bodies	A successful evaluation will often depend on the engagement and cooperation of the organisations and individuals involved in delivering the intervention. It will be important to gain their commitment and be clear about what input will be required from them. Often, evaluation fieldwork needs to be built around the delivery body's capacity to take part – and the-time commitment that will be required.
Wider stakeholders	Other stakeholders are often involved, for example, those directly or indirectly affected by the policy. Engagement of this group can be made by inviting them onto the steering group, telling them about the evaluation or including them as participants in the research.
Post-delivery resource	Evaluations typically commence before an intervention has been implemented and continue post-delivery to answer the evaluation questions posed. This can be after the end of the resourcing for the main policy and delivery teams. It is essential that the financial, management and analytical resources are maintained until the end of the evaluation. This ensures that learning is achieved and evidence is synthesised and published to inform in future interventions.

Sufficient time and expertise need to be dedicated to the management of the evaluation to ensure it delivers on time and within budget, is of a sufficient quality, and meets the needs of the policy and delivery team responsible for the intervention.

Dedicating time to oversee and manage the project will also ensure that any issues are raised and dealt with quickly. There will commonly be issues such as changes to the intervention design or execution, or problems with data collection that require speedy changes to the evaluation design. Likewise, emerging findings can change the focus of interest and require changes to the evaluation design. Close management and collaboration ensure that data collection remains relevant to the learning needs of those running the evaluation and maximises the impact of useful findings. The evaluation manager does not necessarily need to be an evaluation specialist (although this is more efficient) but specialist input will be required.

5.3 Governance

It is crucial to set up effective governance structures at the beginning of an evaluation. This will help in the scoping stage (Chapter 2), commissioning the evaluation, and steering the evaluation through to delivery. It is particularly important where a more flexible evaluation design is required, and when evaluation plans need to evolve over time. It can also be advantageous for key stakeholders to be aware and bought into an evaluation, and to have early sight of its emerging findings.

Typically, evaluation is designed and managed by analytical specialists: evaluators, social researchers, economists, operational researchers or statisticians. Evaluations are most successful if the design is closely aligned with the needs of stakeholders who will ultimately use the findings (see section 1.10).

With agreed governance processes in place, the roles, responsibilities, expectations and potential conflicts between stakeholders can be clarified and managed.

Governance arrangements usually involve the following:

1. **Policy programme/project board:** For evaluations of government interventions, buy-in from decision-makers, and their use of the findings, is essential. This can be achieved through ensuring the policy team formally owns the evaluation and that the evaluation team sit on/regularly report progress to the policy board. This can allow the policy board to:

   - Help shape and steer the direction of the evaluation.
   - Agree resources.
   - Clearly understand what the evaluation will (and will not) deliver.
   - Ensure the evaluation adapts to changing circumstances (such as the dates of key decision points).
   - Be aware of emerging findings.
   - Consider alignment and consistency between project delivery/benefits management reporting and evaluation findings.
   - Use emerging findings.

2. **Evaluation steering group:** A steering group is focused solely on ensuring the evaluation meets its objectives and responds appropriately to emerging issues. It should include all the main stakeholder voices: the evaluators, the commissioners and representatives of the intervention policy team. Ideally, it should also include those responsible for the intervention delivery, participants and, for larger projects, independent expert scrutiny and advice.

The steering group should meet throughout the evaluation, from design stage onwards,[72] continuously communicating and engaging with the evaluation team, so that it can genuinely steer progress and address any unforeseen issues. Its tasks are to:

- Advise on the evaluation design and whether it is likely to meet the intended goals of the evaluation.
- Consider whether the methods proposed are feasible.
- Provide guidance to ensure delivery of a high quality and policy-relevant evaluation.
- Provide advice on how to proceed if circumstances change.
- Facilitate the work of external evaluators.
- Provide access to information and contacts.
- Quality assure the evaluation design, questions, methods and research tools.
- Advise and quality assure the analysis and interpretation of the evidence.
- Provide oversight on the appropriate use of AI tools, ensuring the evaluation's quality is not diminished and that the ethical implications have been properly considered.

3. **Expert peer review:** Expert peer review is useful at all stages of an evaluation although it is most common at the design and reporting stages. Peer review allows experts independent of the intervention and its evaluation to assess whether the questions, design, execution and findings are fit for purpose and meet the stated objectives. It also allows an assessment of and compliance with ethical, legal and commercial processes. Peer review can be conducted internally by individuals unconnected to the evaluation, or externally by evaluation or subject experts.

   External peer review is typically conducted by notable experts in the evaluation approach or method used, or in the subject area studied. They offer high-quality scrutiny and challenge, and their expertise and independence can enhance the credibility of the evaluation and bring a wider range of expertise and points of view. Care needs to be taken when seeking external advice to avoid conflicts of interest.

72 A steering group should ideally be formed before the evaluation is designed, to ensure all stakeholders are able to contribute and agree the overarching evaluation questions, scope and design. A steering group member who does not agree with the fundamental elements of the evaluation is unlikely to be content with the evaluation as it progresses.

5.4 Linking evaluation to the intervention design

As described in Chapter 2, the evidence requirements from the evaluation can usefully influence the design of the intervention to enhance the quality of the evidence generated and the resulting learning that can be achieved. The plan for the evaluation should, therefore, be designed alongside that of the intervention.

Key ways the intervention can support the evaluation include:

1. **Building evaluation data requirements into monitoring data collection:** Simple adjustments to the collection of monitoring data can provide rich data that can be used by the evaluation. This is a cost-effective way of collecting information and can make a survey of participants either unnecessary or shorter.
2. **Collecting contact details and consent:** Alongside informed consent from participants agreeing to participate in the research, it is also necessary to get consent for any proposed data linking or for contact details to be held and used for follow-up research (see section 5.9 on Ethics). In such cases, you should get advice from GDPR experts (see section 4.5, Data handling).
3. **Conducting a census of participants:** A cost-effective way to collect data can be through a census or automatic survey of all participants. This could mean that:
   - All participants are contacted and invited to take part in the survey.
   - Taking part in the survey is a mandatory requirement of the intervention (for example, applicants for a grant are required to complete a survey). In this case, the requirement needs to be built into the intervention and clearly explained in any advertising.
4. **Piloting/testing:** Piloting and testing allow an intervention to be tried out on a smaller scale to understand (a) if it works (b) whether it can be modified to improve it and (c) how it is best delivered. This is essential if there is any uncertainty around the likely success of an intervention.

5. **Targeted implementation to create comparison groups:** As discussed in Chapter 3, some evaluation methods require both an intervention and comparison group to measure the impact of the intervention. The method of assigning individuals or groups to the intervention or comparison group is best built into the intervention design itself. Two of the most common ways to achieve this are:

    - **Randomised control trials (RCTs).** Randomly assigning individuals or groups to the treatment or control group. This needs to be done as part of the intervention design.
    - **Phased implementation.** One group is subject to the intervention before another to allow for a comparison between the two groups. This approach can work if it is expected that outcomes for the first group can be observed in the short term, before the second group receives the intervention.

It should be borne in mind that some data access or data collection can require legislation – for example, the automatic collection of specific details about participants in a national scheme. If this is the case, the data requirements need to be identified early and built into the intervention's legislation. Without such data, identifying participants and obtaining data from them can be very difficult.

5.5 Specifying an evaluation

Evaluations can be conducted in-house by departmental analytical specialists or contracted out to specialist teams. Even when conducted in-house, it is often necessary to externally commission certain parts of the evaluation (such as collection of new primary data). The decision as to which route to take will depend on:

- The capacity and capability of the in-house analytical team. Evaluations can be resource intensive, especially if research fieldwork is involved, and can use a number of sophisticated techniques that require expertise, as well as specialised software or hardware.
- The need to demonstrate independence. This can improve the credibility and trust put in the findings.
- The timeframes involved. Evaluations can often be long-term projects and internal resources may change. Over-reliance on specific individuals can cause problems.

Regardless of the commissioning route, a robust specification is essential. Specifications can be very tight, setting out exactly what is expected to happen and when – or loose, setting out the work and timescale objectives, but asking those bidding for the work to suggest appropriate methods. There are pros and cons to both approaches. The advantages of a very tight specification are that the commissioner knows exactly what to expect, can set standards and can compare bids based on price, as the same methods will be proposed in each bid. The disadvantages are that it can miss creative ideas from external experts and might discourage those who are keen to 'problem solve' and be innovative in their evaluation design.

Typically, specifications take a middle ground by specifying the evaluation questions to be answered, the suggested approach to be taken and the kind of methods that might be used, but leaving bidders open to propose creative solutions.

There is no set practice around whether to specify the budget in the specification of externally commissioned work. However, a broad price range can be useful in loosely specified work to help bidders assess the scale of the work and ensure that bids do not come in over budget.

When specifying an evaluation, it is important to include:

- The purpose and intended use of the evaluation evidence.
- The timing of any decisions that will be informed by the evaluation.
- The recommended evaluation approach and/or methods where these exist.
- The data that already exist (the intervention's own monitoring data or other available and relevant data).
- Any specific essential evaluation activities.
- An indicative publication strategy that sets out what will be published and when.
- The proposed approach to integrating AI (where appropriate) in the evaluation design, evaluation delivery or both, and considering and handling any associated risks with AI use.

It is useful to request bidders to provide a synthesis strategy: the plan for how the various research and data streams from the evaluation will be brought together into a clear narrative that answers the evaluation questions (see Chapter 3).

The quality of the specification of the work will determine the quality of the ultimate evaluation. Time and expertise dedicated to preparing a well-thought-through, clear, precise specification will be well spent. For complex evaluations, more information on specifications for commissioning can be found in the Supplementary Guidance on 'Evaluating Complexity'.

5.6 Utilising Artificial Intelligence (AI) in evaluation

AI systems offer an opportunity for evaluators to accelerate their work but they also come with risks. AI is a wide range of techniques, including machine learning, natural language processing and generative AI, to mimic human intellect and perform tasks such as problem-solving, understanding and generating human text, and decision-making.

AI tools can reduce the time it takes to complete a task that would have taken a human alone much longer, reducing the cost of the evaluation. It can reduce or even eliminate mundane tasks, freeing time for more creative, higher value work. It may help raise the quality of an evaluation if managed well. At the same time, there are new challenges and risks associated with this emerging technology. The use of AI should be proportionate, quality-assured and verifiable.

The UK government has highlighted general principles and safeguards that should be taken into account for safe and responsible AI usage in the Generative AI framework for HM Government[73] and the AI Playbook for the UK Government.[74]

Understanding AI helps evaluators decide whether and when it is appropriate to employ it in their workflows. AI operates on a probabilistic basis. Rather than following rigid rules, AI systems are trained on extensive data to identify patterns. When given an input, an AI model calculates the probability of various potential outcomes and usually selects the most likely one. In practice, this means that AI is making an 'educated guess' based on the patterns it learned from its training data.

Because of their probabilistic nature, AI models produce output with some randomness (stochasticity), so the same input may not reproduce the same output each time. Variation in outputs and replicability of AI-generated evidence or findings over time can also result from AI models being updated, often without notice. At the same time, AI models 'homogenise' – in other words, they produce an 'average' output while reducing the diversity of outcomes.[75]

AI tools can produce output (text, images, statistical analyses, results and so on) that is plausible based on statistical patterns in their training data, but false. It is therefore critical that evaluators remain involved in the process and that they have processes in place to check that AI outputs are accurate and appropriate.

73 HM Government (2023) *Generative AI Framework Report*.
Available at: https://assets.publishing.service.gov.uk/media/65c3b5d628a4a00012d2ba5c/6.8558_CO_Generative_AI_Framework_Report_v7_WEB.pdf (Accessed: April 2026).
74 Government Digital Service (2025) *AI Playbook for the UK Government*.
Available at: https://www.gov.uk/government/publications/ai-playbook-for-the-uk-government (Accessed: April 2026).
75 Doshi, A.R. and Hauser, O.P. (2024) 'Generative AI enhances individual creativity but reduces the collective diversity of novel content', *Science Advances*, 10(28), p. eadn5290.

Many AI systems are considered ‘black boxes’ because their internal processes are not easily understood or explained. These models accept an input and produce an output without showing the steps they took to get there. Additionally, the quality of training data – which is often opaque in proprietary AI models – is paramount: Limited diversity of training data or lack of representation may maintain biases or lead to lack of nuance and diversity in the final output.

Proportionate use of AI should be decided on a case-by-case basis, based on the specific context. Like all tools, AI should be used to complement – not replace – the critical role of evaluators and their expertise, and any risks arising from AI should be actively managed and transparently communicated, alongside associated limitations. The conclusions from an evaluation need to be robust, defensible and meet the quality standards expected of government evaluation. Some examples of different potential uses of AI in evaluation are presented in Table 5.2.

Table 5.2: Examples of the potential use of AI in the evaluation process

Theory of Change	AI may help develop a Theory of Change by, for instance, helping visualise the links between different stages, and by identifying a wider range of potential outcomes, including unintended consequences and long-term impacts.
Review and synthesis of existing evidence	AI tools can accelerate the process of screening a wide range of traditional and non-traditional sources of literature (including academic and grey literature as well as any Internet-based content) and extract of data from various sources. However, the use of AI in evidence reviews also introduces challenges regarding systematic coverage, selection bias and replicability. Lack of access to publications behind paywalls, opaque filtering mechanisms, and resource limits (such as maximum tokens per search) may preclude an exhaustive list of evidence sources, without informing evaluators of the choices being made by the AI. Further, the probabilistic nature of AI may result in different evidence reviews being produced, reducing reproducibility. This is a particular issue for systematic reviews where replicability, comprehensiveness and unbiasedness are critical. Evaluators need to manage these risks and quality-assure the outputs by, for instance, considering human-led screening process against pre-defined inclusion criteria, checking the accuracy (and existence) of sources, and taking steps to increase comprehensiveness systematically.
Experimental and quasi-experimental methods	AI and machine learning techniques can enhance experimental and quasi-experimental designs, for example by informing comparison groups that are matched on a wider set of characteristics or by identifying complex patterns in how an intervention's impact varies across different subgroups. However, even sophisticated AI techniques built on probabilistic associations alone cannot establish causality from data. The methods for impact evaluations covered elsewhere in the Magenta Book remain key resources for establishing causality.
Theory-based impact evaluation methods	AI tools can be used to support theory-based methods by, for instance, gathering evidence from academic literature or policy documents to identify patterns, evidence gaps or alternative explanations that can help build or refine a Theory of Change. However, the evaluator's role in interpreting these patterns and identifying potentially causal pathways remains critical.

Qualitative analysis	The use of AI can significantly accelerate the processing of qualitative data, offering rapid transcription and preliminary thematic analysis of interviews, focus groups or case study notes. However, this efficiency comes with potential risks that evaluators need to manage. AI models can exhibit a 'homogenising' effect, potentially flattening rich narratives, overlooking nuance in coding themes, overlooking subtle contextual cues, or failing to identify minority perspectives that a human analyst would be less likely to overlook. Evaluators need to manage these risks with careful human oversight and quality assurance.
Data analysis	Generative AI tools can be effective and rapid at cleaning datasets as well as generating (and debugging) code for data analysis. Many statistical analysis programmes offer AI-supported coding – from basic 'auto-completion' of specific functions to writing entire sections of analysis code. However, using AI in coding and data analysis without sufficient human oversight can also lead to data security breaches, data not being cleaned correctly, inappropriate statistical models being chosen and output being produced that does not reflect the full range of policy-relevant outcomes. Evaluators need to ensure that their own involvement and oversight at different stages of the process manages these risks effectively.

The following considerations should be kept in mind when using AI in an evaluation:

- **Accountability:** Evaluators need to assume accountability and responsibility for the process and outputs, not delegate these to an AI model. Limited explainability of AI tools and models can undermine confidence in an evaluation and the trust that can be placed in the results. Evaluators should ask themselves whether they sufficiently understand and stand behind how and what AI created for the purposes of the evaluation.
- **Transparency:** A lack of transparency or explainability of new technologies, including AI, can raise concerns of fairness, reliability and accountability of decisions. An understanding of the processes and outcomes, especially those impacting the public, is significant to evaluation. Evaluators should disclose any significant use of AI in evaluations, including which AI tools and which model version (for reproducibility), and acknowledge any limitations that may have resulted from using AI.
- **Reliability and reproducibility:** In stages in an evaluation where exact protocol and reproducibility are essential to reaching a conclusion, evaluators should be cautious: AI's probabilistic nature can introduce variation with unintended consequences (for example, LLMs will produce slightly different output even for identical prompts, which can be problematic for evidence synthesis or qualitative coding). On the other hand, in cases where potential variation is acceptable (for example, how exactly data analysis code is written to produce a figure), AI may help evaluators to be more efficient. Managing this risk is core to the responsible use of AI in evaluation.

- **Handling sensitive data:** As a guiding principle, personal or sensitive data should not be inputted into an AI system unless it is explicitly approved for such use in the Civil Service and adheres to all departmental security and legal protocols. This is because, unless the AI model is officially approved, AI developers may look to use inputs from users to improve their models. This may cause personal or sensitive information to spill into the public domain when an AI model is trained on it in the future.
- **Quality assurance:** Proportionate attention should be spent on quality assuring AI in an evaluation. Quality assuring AI may include verifying the accuracy of all (or, where practically infeasible, at least a representative sample of) outputs to guard against hallucinations, assessing outputs for potential biases inherited from the training data, ensuring AI-produced syntheses include all facts and are reflective of relevant nuances, verifying data used for AI analysis is handled and managed appropriately, documenting the process to ensure transparency and aid replicability, and ensuring that the use of AI has been appropriately disclosed.
- **Evaluator-in-the-loop:** Akin to 'human-in-the-loop' AI development, evaluators should consider establishing a process, by which they are 'in the loop' when AI output is produced and reused at different stages of an evaluation. By inserting human oversight into the process, evaluators can actively monitor and manage risks arising from AI use and respond effectively.
- **Building trust:** Policymakers need to trust that evaluations are a source of reliable evidence. While AI can be an effective tool for specific steps in an evaluation, it cannot engage and build trust with policymakers. Evaluators' comparative advantage relative to AI is to engage with key stakeholders and ensure that both the process and results of an evaluation are trusted.

Similar considerations are also discussed in external guidance for the use of AI in evaluation.[76]

Ethical considerations when using AI in evaluation are covered in section 5.10.3.

Guidance has also been developed to support robust evaluation of AI interventions, see the Magenta Book Annex: Guidance on the Impact Evaluation of AI Interventions.[77]

76 UK Evaluation Society (2025) *AI in evaluation: Good practice guidelines for practitioners*. Available at: https://evaluation.org.uk/new-ai-guidance-launched-for-evaluators/ (Accessed: April 2026).
77 HM Treasury (2025) *Guidance on the Impact Evaluation of AI Interventions*. Available at: https://www.gov.uk/government/publications/the-magenta-book/guidance-on-the-impact-evaluation-of-ai-interventions-html (Accessed: April 2026).

5.7 Commissioning an evaluation

Commissioning an evaluation to external experts is common, especially if new fieldwork is necessary. The precise commissioning route will vary between departments – evaluation project managers should seek advice from departmental commercial experts as to the most appropriate route for them. The commonest routes are via a departmental framework contract – through a cross-government framework, the Crown Commercial Service research marketplace, or through open competition.

There is benefit in maintaining a wide range of potential external evaluators to work with to ensure a constant supply of fresh ideas and to keep competition between suppliers. It can also be useful to invite bidders to collaborate to ensure the full range of skills required for an evaluation are brought together in one bid. 'supplier days', where potential bidders are brought together to hear more about the intervention and the evaluation requirements can be a useful way to encourage interest in an evaluation and to bring together potential consortia partners. When commissioning external suppliers, the specification and subsequent contract should clearly state the expectation that the work will be conducted in accordance with the principles and standards of good practice set out in the Magenta Book.

Assessing bids will be conducted by the project manager and at least two other assessors. These can be members of the steering group, members of the intervention design or delivery team, or other evaluation or analytical experts. The assessment criteria must be clear in the specification published in the invitation to tender and there should be agreement between the assessors of what 'good' looks like before bids are received. Typically, cost information is separated from bids, so that bid assessment is conducted on quality alone. Cost is considered afterwards to differentiate between bids that are deemed to be of sufficient quality.

It is important to note that intellectual property (IP) rights rest with individual bidders. It is not possible to take ideas from one bid and suggest them to another bidder.

In all cases, sufficient time should be given for the commissioning process. The steps involved – gaining clearance for the specification from all stakeholders, gaining procurement sign-off, providing sufficient time for bidders to respond, allowing time for bid assessment and agreement, and awarding contracts – can take several months.

5.8 Flexibility and consistency

There is often a balance to be struck between consistency and flexibility in designing and managing an evaluation. Experimental and quasi-experimental approaches suit consistency. This has clear advantages in enabling comparisons to be made between the intervention and the control group, or between 'before' and 'after' samples but requires both the intervention and the evaluation to remain unaltered throughout delivery. However, interventions can often be complicated and complex. Decision-makers are often keen to tweak and make 'in-flight adjustments' to maximise the anticipated benefits. Likewise, it may be difficult to know in advance which of the evaluation methods used are most likely to yield useful results.

Building in flexibility at the earliest stages will allow the evaluation to adapt to changes in requirements in the intervention design, delivery mechanisms and data collection. This is especially important in evaluations of new or innovative interventions, where the Theory of Change has not been tested and aspects of the intervention design are uncertain. Flexibility will enable evaluators to adapt to respond to emerging findings – to 'dig deeper' or refocus attention – and to respond to the need to report findings at a different point in time.

Communication between the evaluators and stakeholders is key in ensuring that all parties agree changes to the evaluation design and understand its implications. A clear audit trail of changes and decisions must be maintained. Governance processes can help with this.

5.9 Quality assurance

Quality assurance is fundamental to all evaluations. Independent scrutiny is often useful, especially from a range of analytical perspectives, but even without this the project manager must take proportionate steps to ensure the design, execution and findings are of a suitable standard.

Quality assurance is an ongoing task and needs to be planned from the outset. This is to ensure that all aspects of an evaluation – design, execution (including all fieldwork), analysis and reporting – are conducted to appropriate standards and are fit for purpose.

At each stage, quality assurance may be conducted through internal expert/in-house peer review, steering group review or external expert peer review. Often, a mix of these sources is used.

At the analysis stage, quality assurance should check that any analytical procedures have been applied fairly and appropriately. For quantitative analysis, this may include independent reproduction of analysis and review of analysis code.

At the reporting stage, quality assurance ensures the findings are based on an objective, defensible interpretation of the results and relate to the original objectives of the evaluation. The limitations and caveats of the analysis should be clearly communicated, and inclusion of a technical methods annex can also strengthen confidence and aid replicability.

Wherever possible, all reported numbers should be checked. Further checks might include observing data collection (for example, listening to telephone interviews or observing focus groups), reviewing the analytical framework, or checking a sample of analytical results to ensure they can be reproduced from the raw data using the documentation provided by the contractor. Note that data anonymisation and other ethical considerations may be required here.

5.10 Ethics

Ethical consideration is an active process and should happen throughout evaluation design, delivery and reporting. Evaluation, as with social research, often raises questions of ethics that might influence the evaluation methods, fieldwork and reporting. There are few predetermined 'right' or 'wrong' answers, and decisions often involve weighing up competing obligations. Table 5.3 below sets out four groups of stakeholders and the ethical considerations associated with each.

Table 5.3: Stakeholders and ethical considerations

Stakeholder	Examples of ethical considerations:
Participants	• Protecting confidentiality. • Avoiding harm (consider where it might be justifiable to break confidentiality because of concerns over respondents' wellbeing). • Minimising respondent burden and avoiding intrusion. • Avoiding manipulation or deception.
Colleague/ partner agencies	• Reporting of controversial or potentially damaging findings (such as reputational). • Additional burden created through collection and sharing of data. • Authorship and appropriate credit.
Funders, employers and researchers	• Tendering rules and procedures. • Contractual clarity and division of responsibilities. • Rules/norms of publication. • Protecting fieldworkers. • Whistleblowing.
Wider society	• Protecting the vulnerable. • Publishing publicly funded research. • Being honest about the limitations of research.

5.10.1. Ethical principles

There are a number of sources of guidance on research ethics, including the Social Research Association,[78] the Economic and Social Research Council[79] and the UK Evaluation Society.[80] In government, the Government Social Research (GSR) service has guidance on ethics,[81] which is applicable to evaluation. It has five principles:

1. **Sound application and conduct of social research methods, and interpretation of the findings**

   All methods, analysis and reporting should be fit for purpose and able to withstand robust external scrutiny and meet a real unmet need.

2. **Participation based on informed consent**

   Informed consent must be obtained from research participants to allow the collection, analysis, transfer, storage and linking of data. Informed consent is an ongoing agreement by a person to participate in the evaluation on the basis that they understand:

   - The purpose of the research.
   - Their role in the study.
   - How data about them will be managed.
   - How data about them will be used in the future.
   - That their participation is voluntary/they can withdraw at any time.

   Informed consent raises a number of legal issues that are outlined in:

   - The Data Protection Act (2018) (DPA),[82] which sets out regulations for the processing of information relating to individuals, and what information individuals can request about themselves.
   - The General Data Protection Regulations (2018) (GDPR),[83] which set out rules on controlling and processing personally identifiable information.
   - The Equality Act (2010)[84], which requires public bodies to ensure their work supports equality by treating people from different groups fairly and equally. Therefore, evaluations should be conducted in a way that enables people from different groups to participate.
   - The Mental Health Act[85], which stipulates that research should only involve those lacking mental capacity under certain conditions.
   - The Freedom of Information Act (2000)[86], which provides a right of access to information held by a public authority, including research information.

78 Social Research Association (n.d.) *The Social Research Association Official Website*.
Available at: http://the-sra.org.uk/ (Accessed: April 2026).
79 Economic and Social Research Council (ESRC) (2019) *Economic and Social Research Council Official Website*.
Available at: https://esrc.ukri.org/ (Accessed: April 2026).
80 UK Evaluation Society (2019) *Good Practice Guidelines*.
Available at: https://www.evaluation.org.uk/professional-development/good-practice-guideline/ (Accessed: April 2026).
81 Civil Service Government Social Research Unit (2011) *GSR Professional Guidance – Ethical Assurance for Social Research in Government*.
Available at: https://www.gov.uk/government/publications/ethical-assurance-guidance-for-social-research-in-government
(Accessed: April 2026).
82 GOV.UK (2018) *Data Protection Act 2018*.
Available at: https://www.gov.uk/government/collections/data-protection-act-2018 (Accessed: April 2026).
83 Information Commissioner's Office (2018) *Guide to the General Data Protection Regulation*.
Available at: https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/ (Accessed: April 2026).
84 GOV.UK (2013) *Equality Act 2010 Guidance*. Available at: https://www.gov.uk/guidance/equality-act-2010-guidance (Accessed: April 2026).
85 Legislation.gov.uk (2007) *Mental Health Act 2007*.
Available at: https://www.legislation.gov.uk/ukpga/2007/12/contents (Accessed: April 2026).
86 Legislation.gov.uk (2000) *Freedom of Information Act 2000*.
Available at: https://www.legislation.gov.uk/ukpga/2000/36/contents (Accessed: April 2026).

3. Enabling participation

It is against the law to discriminate against anyone on the basis of 'protected characteristics', such as age, disability, sex, sexual orientation or race (see Public Sector Equality Duty).[87] Potential barriers to consider include geographical, cultural, financial and communication. The evaluation should consider issues likely to act as a barrier to participation (such as people with disabilities, people living in excluded communities where interviewers are unwilling to travel, people whose first language is not English, or people without a permanent address) and outline reasonable steps taken to address them.

The pros and cons of taking steps to reach potentially excluded sections of the population need to be weighed up for each individual research tool, and inclusion strategies should be developed at the design stage.

4. Avoidance of personal and social harm

Individual research participants (including those who drop out), the wider social groups or organisations to which they belong, and researchers should have their physical, social and psychological wellbeing protected at all stages of the research process.

Intrusion caused by the collection of data for an evaluation should be avoided and the privacy of respondents should be respected. In some cases, participants may find the process of sharing information upsetting or traumatic (for example, some people may find talking about illness and/or loss of a job difficult).

GSR recommends that where there is 'more than minimal' risk to participants, a formal risk assessment may be appropriate, especially where research involves vulnerable groups (like children, offenders or disabled people) and/or deals with a socially sensitive issue, such as mental health.

5. Non-disclosure of identity

The identity of, and data belonging to, participants and potential participants (including information about the decision whether or not to participate), should be protected throughout the research process, including respondent recruitment, data collection, data storage, analysis and reporting.

5.10.2. Ethics in social media research

The use of social media in research and evaluation is still developing, and ethical principles cannot always be applied in the same way as in more traditional research methods. The government's Social Media Research Group has developed guidelines[88] based on the GSR's five key principles, set out in Table 5.4.

87 GOV.UK (2012) *Public Sector Equality Duty*.
Available at: https://www.gov.uk/government/publications/public-sector-equality-duty (Accessed: April 2026).
88 Social Media Research Group (2016) *Using Social Media for Social Research: An Introduction*. Government Social Research.
Available at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/524750/GSR_Social_Media_Research_Guidance_-_Using_social_media_for_social_research.pdf (Accessed: April 2026).

Table 5.4: Ethical principles for social media research

Principle	Key considerations:
1: Sound application and conduct of social research methods and interpretation of the findings	• Are social media techniques the most appropriate method? • Are methods being used professionally? • Is appropriate quality assurance in place? • Are details of the project publicly available, including the research purpose and the data being used?
2: Participation based on informed consent	• Do the terms and conditions users have signed up to cover the collection, analysis and use of their data? • If individual informed consent is required, how will participants be contacted? • If data is posted to a social media platform and then deleted does previously given consent remain valid?
3: Enabling participation	• Are any groups being inappropriately excluded given the nature of the research questions and/or platforms used?
4: Avoidance of personal and social harm	• Are the data public or private? Any research involving private content should only be conducted with explicit informed consent from the user. • How will the collection of unnecessary personal data be minimised?
5: Non-disclosure of identity	• How will the identity of users be protected? (There can be no guarantee of full anonymity with social media research. If researchers wish to include verbatim content, they should consider contacting social media users to ask them if they would be happy for their content to be cited.)

5.10.3. Ethics in AI-assisted evaluation

The use of AI in evaluation introduces ethical duties. While AI can be a useful tool to speed up or improve aspects of the evaluation, evaluators remain fully accountable for the process and its outputs. This professional responsibility cannot be delegated to an AI model, and evaluators must be able to defend the methodology and conclusions, understanding the tool's limitations and potential for error.

For instance, a key duty is ensuring algorithmic fairness, as AI can amplify biases present in its training data, and evaluators must therefore proactively mitigate the risk of generating inequitable outcomes. It is also important that evaluators understand the implications of the probabilistic and correlational nature of AI on the output it produces, and evaluators must ensure they do not arrive at conclusions that cannot be replicated. See also the practical principles of using AI in evaluation in section 5.6.

Furthermore, if AI is used during the data collection or analysis process, informed consent must be meaningful. Participants should be made aware that their data will be processed by automated systems, upholding the principle of treating their contributions with respect. It should be noted that personal or sensitive data should not be inputted into an AI system unless it is explicitly approved for such use in the Civil Service and adheres to all departmental security and legal protocols.

To maintain public trust, evaluators must be transparent about where and how AI has been used. Ethical oversight of AI is not a one-off task but a continuous responsibility for governance bodies throughout the evaluation life cycle.

HM Treasury

Chapter Six

The use and dissemination of evaluation findings

6

Summary

The value of an evaluation comes through its use and influence. Planning for this in advance will enhance the usability and use of evaluation outputs.

An explicit use and dissemination plan will ensure that all users and uses are captured, and that the outputs will be framed to maximise impact with these groups.

Regular reporting can ensure the findings are available to use in decision-making throughout the course of an evaluation.

The GSR Publication Protocol is a useful framework to use when considering the publication of findings.

There should be a presumption towards openness for all evaluation findings and materials. This includes underlying data and research instruments.

6.1 Introduction

An evaluation's value comes through its use and influence. This should be considered at the planning stage and continually revisited throughout the evaluation's execution.

The users of an evaluation can be direct or indirect.

- **Direct:** those designing and implementing an intervention use the findings to improve the intervention's design or implementation and to maximise the chance of achieving the intended outcomes. Those scrutinising government performance use evaluation findings to assess the performance and outcomes of policies.
- **Indirect:** those who use the findings to answer other, related questions, design future policies in the same or similar areas, or capture learning about the use of public funds.

Evaluation outputs need to be carefully designed to meet the needs of the various users. Evaluations with clear and evidenced implications for future decision-making are only of value if the findings are accessible and usable by the relevant decision-makers at the right time. Likewise, scrutiny bodies and the wider public should be able to easily access, understand and use the findings to assess the design, implementation and outcomes of government policies. The challenges of achieving use and influence are common to all organisations seeking to extract maximum value from their research.[89,90]

89 Department for International Development and UK Aid (2014) *DFID Evaluation Strategy 2014 to 2019*.
Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/380435/Evaluation-Strategy-June2014a.pdf (Accessed: April 2026).
90 Nesta (2019) *Using Research Evidence: A Practical Guide*.
Available at: https://www.nesta.org.uk/toolkit/using-research-evidence-practice-guide/ (Accessed: April 2026).

6.2 Developing an evaluation use and dissemination plan

There is no fixed way to develop an evaluation use and dissemination plan. The key is considering the four questions below:

1. **Which groups?** A list of potential stakeholders can be found in section 1.10. It is worth considering all potential users singly as it is easy to miss key stakeholders who might have very different needs from the evaluation. A stakeholder mapping exercise can be useful.
2. **What information?** Knowing what information is needed by each group should inform the evaluation questions.
3. **Which point in time?** There will be decision points throughout the implementation and delivery of any intervention (see Table 2.1). Knowing these points can allow the various data collection plans in place to be designed to meet these requirements. Note: the evidence need often comes before the final decision point to inform thinking.
4. **For what purpose?** Evaluation evidence can be used for various purposes, which will affect the type of evidence generated and how it is used and disseminated. This is particularly important for interim decision points. When the decision is small (for example, a review of staffing numbers to support the intervention), emerging monitoring data may suffice. However, if the decision is large (for example, whether to roll out an intervention nationally), the evidence needs will be much greater and call for a much higher level of robustness.

Identifying these requirements at the evaluation design stage enables discussion, negotiation on timetabling and management of expectations from the outset.

The plan should cover what will be published, when it will be published and which communication tools will be used (such as print publication, hard publication, social media or seminars and conference presentations). It should include the decision points and what evidence will be available when. All outputs to be published should be covered, including the reports, the underlying data and the research tools.

Detailing the 'influence objectives' in the use and dissemination plan can be useful. Influence objectives specify the desired impact of the evaluation. For example, an evaluation could be conducted to identify good practice in prison rehabilitation schemes. The influence objective could be using the evaluation findings to positively change the way that rehabilitation programmes are implemented. Once the influence objectives are clear – evaluation outputs can be tailored to meet this goal.

To gain buy-in, it is good practice to agree the use and dissemination plan with as wide a range of stakeholders as possible; this is typically covered by the programme board and steering group. This early engagement can also help prioritise and manage expectations as it is unlikely that all questions asked can be answered.

Two additional groups that can usefully be approached when developing the use and dissemination plan are departmental press offices (to agree the general principles in the plan and the broad approach to be used) and ministers' offices (to gain their buy-in and agreement to publication plans upfront). This is good practice to avoid the perception that publication decisions are unduly influenced by the nature of the findings.

Once the different audiences and their evidence needs are known and prioritised, reporting and communications should be tailored to meet these needs. The value of tailoring communications this way cannot be underestimated. Improving the usability of the findings by helping specific audiences understand how findings directly relate to their areas of interest can be invaluable in ensuring the findings are used.

Improving the use of the findings involves considering which groups have capacity to act on the findings, in what ways and the constraints they may be under.

6.3 Disseminating evaluation findings

It is essential to identify and use the most appropriate outputs and communication channels to maximise the reach to and impact on the evaluation's various stakeholder groups.

While a final evaluation report is an important document in an evaluation study, it is rarely the best way to ensure impact and influence. For example, front-line delivery staff and ministers will have very different information needs and preferred communication styles. Being able to input emerging findings into programme board discussions can be invaluable to the group's deliberations.

Alternative modes of communication include one-page summaries, video outputs, infographics, data sharing, newsletters, social media posts, conference presentations and seminars. Discussing evidence needs with stakeholders can help identify the most effective communication channels for that group. It should be noted that it is not appropriate to discuss findings in public before they have been published in some form. However, ensuring stakeholders, such as steering groups and policy/delivery leads, are on board with the messages prior to publication can be useful, particularly in the handling of negative findings.

AI can be useful for generating tailored summaries, presentations or data visualisations for different stakeholder groups. While this can increase the speed and reach of dissemination, all AI-generated content must be rigorously checked by evaluators for accuracy, tone, and the correct representation of the evaluation's conclusions and its limitations.

Publication can be resource-intensive for government evaluation managers, as it requires quality assurance, editing, ministerial clearance and press office input. Where appropriate, an option can be to publish regular, short outputs to ensure findings are released and can be openly used in a timely fashion.

Academic contractors may seek to publish evaluation outputs in academic journals, with the advantage of exposing the work to more people, providing further quality assurance and spreading knowledge. Encouraging this at the start of a project can also attract a wider range of academics to take part in evaluation activities.

6.4 Building an evaluation culture

To understand evaluation as an embedded practice means that it should not operate as a separate, independent function but more as an integrated part of an organisation's culture and operational structure. This means that evaluative and reflective practice is part of the 'way we do things around here' where all colleagues seek, learn and think critically about the evidence that underpins their actions.

Embedding this way of thinking rarely happens by accident. It implies a shift away from seeing evaluation as a necessary evil or a ritualistic necessity, to one in which attributing value to the implementation of an intervention is built in as a core dimension of good policy practice. It strongly implies 'using' evaluations to guide future actions and 'next steps' and is particularly relevant to working in complex environments (see the Supplementary Guidance on 'Evaluating Complexity').

In day-to-day terms, it means department heads or specific initiatives need to emphasise the expectation that an evaluative dimension is required at the policy design stage. Colleagues should be actively encouraged to consider how the value of a policy will be established as a matter of course.

6.5 Publication

The GSR Publication Protocol[91] provides cross-government advice on the publication of research, including evaluation. It covers all aspects of publication through the following five principles.

- **Principle 1:** The products of government social research and analysis will be made publicly available.

  This establishes a presumption that all outputs will be published.

- **Principle 2:** There will be prompt release of all government social research and analysis.

  This sets a 12-week maximum from agreement of a final output to publication.

- **Principle 3:** Research and analysis must be released in a way that promotes public trust.

  This establishes the features of a robust study, clear of ministerial interference.

- **Principle 4:** Clear communication plans should be developed for all social research and analysis produced by government.

  This suggests that departments should publicly announce what research projects have been commissioned and that communication plans should be drawn up for all projects as part of project management principles.

- **Principle 5:** Responsibility for the release of research and analysis produced by government must be clear.

  There should be a named person in each department who is responsible for ensuring the protocol is adhered to.

91 Government Social Research (2015) *Publishing Research and Analysis in Government: Government Social Research Protocol*. Available at: https://www.gov.uk/government/publications/government-social-research-publication-protocols (Accessed: April 2026).

6.6 Openness and transparency

As suggested in section 1.13, openness and transparency are central to producing useful, credible, robust evaluation evidence. Importantly, openness applies across the whole evaluation process, including in the planning and design stages, in the analysis of data and in the reporting of findings.

In the **planning and design phase**, evaluators should consider how to communicate evaluation plans openly and transparently. Registering evaluations on the Evaluation Registry (a requirement for evaluation reports produced by central Central Government Departments and evaluation of Government Major projects, including those delivered by external organisations) ensures that there is a record of approved evaluations. Additional steps such as publishing evaluation protocols and statistical analysis plans also represent good practice. Pre-registration (see section 2.3.3) of this type provides a transparent record of what the evaluation plans are before data collection or analysis begins, thereby ensuring decisions on the evaluation design are not dependent on the results they produce.

In the **analysis phase**, evaluators should ensure that transparency also applies to the way evaluations are conducted. For larger, more resource-intensive or higher stakes evaluations, where it is feasible and proportionate to do so, evaluators should make evaluation materials, code and data open and accessible so that others are able to understand exactly how findings were produced. When materials, code and data cannot be shared publicly, departments should preserve exact copies internally and provide clear documentation so that it is possible for departments to reproduce and build upon evaluation findings in the future. It is important that any data is stored in a persistent fashion so that it is not lost when individuals or teams involved in the evaluation change. Openness at the analysis stage strengthens the reproducibility of evaluations, allows others to learn from analytical approaches, and increases the credibility and value for money of evaluation investments by making it easier to build on prior work.

In the reporting stage, evaluation outputs should provide a clear account of what was done, why it was done, and how the results should be interpreted. Publishing technical annexes, data tables or peer review comments gives the detail essential for effective scrutiny by specialists. Clear communication of uncertainty is also important so that findings are not presented as more definitive than the evidence allows. Transparent reporting ensures that evaluations are not only credible in the moment but also provide a reliable resource for future learning.

By embedding openness throughout the evaluation process, departments can create a stronger evidence base, enable constructive challenge, and ensure that evaluations deliver maximum public value. For more detailed guidance on these topics, readers should consult the accompanying Transparency in Government Evaluation Research (TIGER) annex.

HM Treasury

Chapter Seven

Evaluation capabilities

7

Summary

Evaluation managers need to have capabilities in four areas: scoping, leading and managing, methods, and use and dissemination.

Scoping capabilities include understanding the rationale for evaluation; constructing an intervention's Theory of Change; identifying the right evaluation approach; and producing a proportionate evaluation plan.

Leading and managing capabilities include demonstrating leadership to maintain momentum and impact; working collaboratively and influencing stakeholders; adapting to changing circumstances; and displaying integrity and resilience.

Methods capabilities include awareness of how to use monitoring and administrative data; methods of primary research and analysis; and methods for impact evaluation (experimental, quasi-experimental and theory-based), process evaluation, value for money evaluation and research synthesis.

Use and dissemination capabilities include reporting and presenting data; considering policy implications; and disseminating evaluation evidence.

An Evaluation Capabilities Framework provides further information (see Supplementary Guidance, 'Government Analytical Evaluation Capabilities Framework'). A self-assessment tool is available for evaluation managers to use.

7.1 Introduction

Drawing on previous chapters, this chapter brings together and summarises the knowledge and skills required by evaluation managers to design and deliver evaluations in government.

The focus of this chapter (and accompanying annexes) is on the government evaluation project manager, typically an analyst from one of the main analytical professions, and the knowledge and skills they are likely to need. An Evaluation Capabilities Framework (see Supplementary Guidance, 'Government Analytical Evaluation Capabilities Framework') provides further information and contains a self-assessment tool for an evaluation manager to use.

7.2 Scoping

At the start of the evaluation journey the evaluation manager will need to:

1. **Understand and communicate the rationale for evaluation:** Evaluation managers need to be able to clarify with stakeholders the purpose and requirements of the evaluation. They also need to be able to champion the value of the evaluation and how it can support accountability, measuring impact, learning and programme development.
2. **Understand the intervention and construct a Theory of Change:** Working with stakeholders, the evaluation manager needs to be able to develop the Theory of Change, linking factors within the evaluation to observable outcomes. They need to ensure the evaluation and the theoretical framework captures and aligns with the programme objectives in the business case, the approach to benefits mapping for the intervention and the assumptions in the appraisal.
3. **Identify the appropriate evaluation type and approach:** The evaluation manager needs to understand the range of evaluation approaches that can be used and determine the most appropriate approach and method for a particular situation. They need to be able to devise appropriate evaluation questions – refined to meet policy needs – and assess the extent that the questions can be answered reliably and credibly (assessing evaluability).[92] Where relevant, they also need to ensure arrangements are in place to access and use monitoring data for the scheme.
4. **Produce a proportionate and appropriate evaluation plan:** Evaluation managers need to be able to design evaluations for in-house work or commissioned projects and to procure external evaluations where appropriate. As part of this, they need to understand the trade-offs between timeframes for observing impacts and policy timeframes. They also need to ensure that the aims and objectives of evaluation work are considered alongside available resources to develop realistic and deliverable proposals.

92 Davies, R. (2015) *Evaluability Assessment*. BetterEvaluation.
Available at: http://betterevaluation.org/themes/evaluability_assessment (Accessed: April 2026).

7.3 Leading and managing

Delivering evaluations – often in challenging and changing policy contexts – requires strong leadership and collaboration skills. The evaluation manager will need to show these skills and behaviours in the following ways:

- Demonstrate leadership and champion evaluation to maintain momentum and impact.
- Work collaboratively, manage relationships and support others (such as contractors and colleagues).
- Proactively influence stakeholders and ensure buy-in from less engaged evaluation users.
- Be aware of and adapt to changing circumstances, feeding this into evaluation delivery.
- Display integrity and resilience and maintain standards when challenged.
- Build a culture of evaluation, learning and transparency into the policy and analytical teams they work in.
- Work effectively across analytical disciplines.
- Be aware of key departmental procurement/commissioning procedures.
- Identify and set up appropriate governance structures for input and quality assurance of evaluation products.
- Ensure all parties are fully aware of their responsibilities and key goals/deadlines are met.
- Consider and implement relevant government guidance that impacts on evaluations (including equality, transparency and ethics).
- Be prepared to invest time and energy in personal development and be willing to work with new methodologies to meet particular challenges.

In addition to these management characteristics, recent writing on new directions in evaluation underscores the need to encourage 'Evaluative Thinking'. Vo et al,[93] in their review of the literature, list some of the key features of Evaluative Thinking. In particular:

- Investigation of one's own (the evaluator's) as well as others' assumptions, motivations and biases.
- Understanding the relevance of context dependency.
- The ability to navigate uncertainty (of outcomes, for example), ambiguity (in data and evidence, for example) and complexity.

93 Vo, A., Schreiber, J. and Martin, A. (2018) 'Toward a Conceptual Understanding of Evaluative Thinking', in Vo, A. and Archibald, T. (eds.) *Evaluative Thinking. New Directions for Evaluation*, 158. San Francisco: Jossey-Bass and the American Evaluation Association, pp. 29–47.

7.4 Methods

The effective design, management and delivery of evaluations requires a range of data collection and analysis skills. Individuals from different analytical disciplines may have strengths in different areas, and the evaluation manager may want to bring other experts in to complement their strengths as the evaluation requires.

The data collection and evaluation methods required are detailed in Chapter 3 and Chapter 4. In summary, the requirements for a government evaluation project manager are:

1. **Use of monitoring and administrative data:** To understand the benefits of using monitoring data in an evaluation, when and how to use it. This includes an awareness of the legal, ethical and practical issues in ensuring programme monitoring data is collected and used effectively. It requires the ability to identify sources and weigh up strengths and weaknesses of data and key metrics to help answer key evaluation questions, especially at the design stage.
2. **Methods of primary research and analysis:** A key skill required is a knowledge of qualitative and quantitative data collection and data analysis methodologies, and the ability to demonstrate practical application of these (for example, knowing how to design and apply specific methods). An evaluation project manager also needs to ensure the necessary quality assurance processes are built into data collection and analysis methods at appropriate stages.
3. **Process evaluation:** It is important to understand the value and role of process evaluation, and its link to monitoring and to impact evaluation, to understand how a programme is working and why. Project managers need to establish key process evaluation questions from the Theory of Change and identify stakeholder groups from which to collect data. They need to be able to time process evaluation deliverables to coincide with opportunities to have impact, and to identify and create suitable interim evaluation products.
4. **Theory-based approach to impact evaluation:** An evaluation project manager needs to be familiar with theory-based evaluation and theory-based methods, their appropriateness for addressing different evaluation questions, and when to use a theory-based approach. They need to be able to identify and use appropriate tests and techniques to assess causation.
5. **Experimental approach to impact evaluation:** An evaluation project manager needs to appreciate the importance of counterfactual and control groups in determining attribution and be able to apply experimental or quasi-experimental methods to attributing impacts. They need to understand when and how to apply RCTs and what robust delivery looks like depending on the assumptions and conditions required for each method.

6. **Value for money evaluation:** This requires a good understanding of the HMT Green Book guidance on economic appraisal and cost-benefit analysis.[94] It is also important that the evaluation manager can use economic techniques to help improve assumptions and methods in any cost-benefit analysis being developed for other impact assessments/appraisals.
7. **Research synthesis:** To bring evaluation evidence together effectively and clearly requires an evaluation project manager to understand a range of methods for synthesis and meta-analysis and the role of triangulation of evidence.
8. **AI-assisted analysis:** An evaluation project manager should understand the potential and limitations of using artificial intelligence (AI) as a suite of analytical techniques. While AI excels at rapidly identifying complex patterns and correlations in large datasets, they must recognise that these methods are probabilistic and cannot replace rigorous causal inference. A critical capability is the ability to judge when the use of AI is proportionate and appropriate for a given evaluation question, and when traditional methods are required to ensure robustness and transparency.

Further detail can be found in the Supplementary Guidance on 'Capabilities'.

94 HM Treasury (2026) *The Green Book: Central Government Guidance on Appraisal and Evaluation*. London: Crown Copyright. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/685903/The_Green_Book.pdf (Accessed: April 2026).

7.5 Use and dissemination

Skills required to deliver fit-for-purpose reporting, use and dissemination include:

1. **Reporting and presenting data:** An evaluation project manager needs knowledge of reporting formats and styles. This includes staying up to date with new and innovative approaches to be able to report evaluation findings clearly and effectively, both orally and in writing. They need to make use of regular reporting where possible to enable 'real-time learning'. They also need to be clear about any limits of the evaluation and to be able to communicate these.
2. **Considering policy implications and sharing the findings:** An evaluation manager will need to understand the policy landscape and build relationships, identifying stakeholders and decision points in order to maximise the value and use of the findings. This means knowing which groups require what information at which point in time and for what purpose and developing a use and dissemination plan from the start to meet these needs.

The evaluation manager needs to be able to tailor the reporting and communication of findings to meet the needs of different audiences. Demonstrating how the findings relate to specific audiences' areas of interest can improve the usability of findings. Knowledge of a variety of communication modes and tools will allow them to maximise impact (see Chapter 6).

Evaluation managers need to be able to take forward publications of outputs beyond the report, including datasets, technical annexes and research tools.

7.6 Further detail

The above list of skills and knowledge are a summary of what an evaluation manager will need. A more detailed list can be found in the Supplementary Guidance, 'Government Analytical Evaluation Capabilities Framework', which provides a fuller description of these skills and knowledge. A self-assessment tool is also available for evaluation managers to assess themselves against.

The UK Evaluation Society also has a 'Framework of Evaluation Capabilities'.[95]

95 UK Evaluation Society (2012) *Framework of Evaluation Capabilities*.
Available at: https://www.evaluation.org.uk/professional-development/framework-of-evaluation-capabilities/ (Accessed: April 2026).