HM Treasury

May/2026

Test and Learn

Magenta Book Annex

This annex was written by:

Jo Voisey, Prototyping Lead in the Evaluation and Prototyping Hub, Ministry of Justice Analytical Directorate

Gloria Crabolu, Senior Lecturer in Systems Thinking and Sustainability, University of Exeter Business School, and British Academy Policy Fellow, Evaluation Task Force, Cabinet Office

Jessica Hunt, Head of Product and Strategy, Test Learn Grow, Cabinet Office

Moria Sloan, Deputy Director, Public Sector Innovation, Cabinet Office

The authors are grateful to colleagues across government and to academic experts who reviewed and commented on earlier drafts of this annex. Their considered feedback and professional challenge have been invaluable in improving the quality and rigour of the final version.

Table of Contents

Foreword

Across government there is growing recognition that the way the state designs and delivers policy must change. The scale and complexity of the challenges facing public services, from improving health and justice outcomes to supporting families, tackling homelessness and delivering the transition to net zero, means that traditional approaches to policy design are often no longer sufficient. The public rightly expect services that work in practice, respond to their needs, and deliver value for money.

Meeting this challenge requires a shift in how government learns.

Traditionally, policies and programmes have often been designed in detail before implementation, with evaluation taking place after substantial investment and rollout. While rigorous evaluation remains essential, this approach can mean that key assumptions about delivery, behaviour and system dynamics are not tested until much later in the policy cycle. When policies are put into practice, it can become difficult and costly to adapt them.

Particularly in complex systems, policy outcomes rarely result from a single intervention. In practice, policies operate within wider systems made up of different organisations, services, and people. Outcomes are shaped by how these parts interact in the real world and might be different when applied in different places and contexts. This means that the effects of a policy cannot always be predicted in advance, and important insights often only emerge once ideas are explored and tested in practice.

Test and Learn provides one practical way to address these challenges. It encourages teams to build a shared understanding of the problem and the wider system, identify and test critical assumptions early, generate evidence through real-world experimentation, and refine policies and services through iterative feedback. By starting small, learning quickly, and adapting based on evidence, government can improve the design of interventions before committing to large-scale implementation and evaluation.

Test and Learn should not be seen solely as a method for analysts or researchers. It is a way of working that brings together policy, delivery, design, partners and analytical expertise to build evidence through action. When embedded effectively, it enables government to learn faster, reduce the risk of costly failure, and improve outcomes for citizens.

In this way, Test and Learn supports the wider ambition of public sector reform: a state that learns continuously, adapts to complexity, and delivers better results for the public.

Beth Russell

Second Permanent Secretary and Head of Darlington Economic Campus at HM Treasury

Glossary

Actor mapping: An actor map is a visual depiction of the key organisations and/or individuals that make up a system, including those directly affected by the system as well as those whose actions influence the system.

Adaptive governance: A flexible way of overseeing projects or programmes, where decisions can be adjusted based on new evidence and learning rather than fixed plans.

Assumptions: Criteria that need to be met for a policy or intervention is to have the intended impact. These may or may not hold true in practice. Test and Learn activity prioritises assumptions that are most uncertain or risky, as these pose the greatest threat to achieving outcomes.

Behaviour diagnosis: An approach to understanding what shapes people's actions, looking at both personal factors (like skills or habits) and wider influences (like rules or social norms).

Beneficiary: The individual, group or community that directly benefits from a policy, programme or service. Beneficiaries' needs and experiences should inform the design, testing and refinement of interventions.

Causal pathway: The step-by-step link between what you do (activities) and the change you expect to see (outcomes).

Causal inference: The process of determining whether an intervention caused a particular outcome, rather than the outcome occurring due to other factors or by chance.

Counterfactual: 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 policy and is commonly called the counterfactual.

Experimentation: A way of generating evidence by deliberately intervening in a policy, programme or service to understand what drives change in practice. It involves intentionally varying one or more elements of an intervention, and systemically measuring the effects. By intervening rather than relying only on observation, experimentation helps teams understand cause and effect, test assumptions, reduce uncertainty and learn which aspects of an intervention should be taken forward.

Feedback loops: A way of learning and improving by checking how things are going, getting people's views or data, and then using that information to make changes. This cycle repeats so the project or service keeps getting better.

Intervention: Anything intended to elicit change, including a programme, policy, project, regulation and changes in delivery method.

Iteration/iterative: Repeating a cycle of testing, feedback and improvement, so that a policy or service gets better, step by step.

Leverage point: A spot in a system where making a change can lead to a big difference in outcomes.

Mechanism: The process or trigger that explains how an activity leads to change – a mechanism can be defined as "capturing people's reasoning and their choices when faced with a policy measure".

Multidisciplinary team: A group of individuals from diverse professional backgrounds (e.g. policy, delivery, design, analysis, behavioural science, different sectors or levels of government) who collaborate to design, test and adapt interventions. Such teams combine different expertise and perspectives to generate better solutions.

Proportionality: In research and evaluation, proportionality means choosing methods whose level of rigour matches the importance of the decision being made, the scale of resources involved, the time available and the potential consequences of getting it wrong.

Prototyping: A simple version or early model of an idea, product or service that is built to test how it works. It can be a rough mock-up (like a sketch or draft) or something tested in a real-life setting on a small, controlled scale. Prototypes let people try things out and give feedback before investing time and significant resources.

Scaling/scale-up: The process of extending a policy, programme or service to a wider population or set of locations, with the aim of achieving similar levels of impact at scale as in the original setting. This requires careful attention to which elements of the intervention are essential for maintaining its effectiveness as it is rolled out.

Service blueprinting: A way of mapping how a service works end to end, showing both what users experience and the behind-the-scenes processes, people and systems that support it. Service blueprints help teams see where things join up, where problems occur, and what needs to change to improve delivery.

Soft systems methodology (SSM): An approach for understanding and improving complex problems where people have different views about what the problem is and how it should be addressed. SSM helps teams to surface and compare different worldviews, make assumptions visible and develop a shared understanding of how the system operates in practice. Rather than seeking a single 'right' solution, it helps people reach agreement on changes that different groups can accept and that can work in the real world.

Stakeholder: Any individual, group or organisation with an interest in or influence over a policy or intervention. This includes policymakers, delivery partners, front line staff, beneficiaries, funders and oversight bodies.

Systems mapping: A visual method for showing how people, organisations, processes and factors connect and influence each other within a system.

Theory of change: A clear description of how a policy or programme is expected to lead to change, showing the steps from activities to outcomes and the assumptions in between. This is typically helped by a visual representation to see the steps in a figure with boxes and arrows showing the causal pathways. See Chapter 2 in the Magenta Book for more information.

Unintended consequences: Outcomes of a policy or programme that were not planned for, which can be either positive or negative.

Waterfall approaches: A way of running projects where each stage happens in order, one after the other (like water flowing down steps). You can't move on until the previous stage is finished, which makes it less flexible if things need to change later.

HM Treasury

Chapter One

Introduction to this guide

1

This annex to the Magenta Book provides a practical introduction to 'Test and Learn' as a way of working[1] for developing and delivering policy. Test and Learn is not a type of evaluation (such as process, impact or economic evaluation) but an approach that integrates research and evaluation thinking throughout the policy cycle to increase the likelihood of achieving the intended outcome/s. Used well, a Test and Learn approach builds the evidence and confidence needed to progress to programme-level evaluation, including the use of counterfactual impact methods where feasible, before decisions are made about scaling up. This annex also explains how the Test and Learn framework has been designed to manage issues of rigour and bias, and to support robust evaluation practice.

Test and Learn

A way of working that uses structured testing, feedback, and adaptation to improve how policies and services are designed and delivered. It focuses on testing critical assumptions early, learning from real-world evidence, and refining interventions before formal evaluation.

This annex complements HM Treasury's Magenta Book and its supplementary guidance on handling complexity, and the two are designed to work together. The Magenta Book provides the overarching framework for high-quality evaluation across government, setting out core principles, standards and methods to determine what works. The Handling Complexity guide extends this by recognising that many modern policy challenges are dynamic, interdependent and unpredictable, and that evaluation therefore needs to be adaptive and system-aware.

The Test and Learn annex builds on both by offering a practical route map for applying these principles in policy and service delivery. It demonstrates how iterative, evidence-informed testing can be embedded throughout the policy cycle, using a blend of social research and experimental methods that are proportionate to the level of risk, investment and potential harm.

1 Refers to the shared methods, behaviours and practices that shape how teams make decisions and collaborate. In a Test and Learn context, it means embedding evidence, feedback and adaptation into everyday practice so that learning is continuous and decisions are grounded in real-world insights.

Test and Learn is most valuable when a problem is complex or uncertain, assumptions are untested, and teams need rapid feedback and small-scale testing to compare different options and adapt to diverse or changing contexts.

Test and Learn is less valuable when the problem and solution are well understood, requirements are fixed or delivery must follow a tightly sequenced, fully planned approach – such as in infrastructure, statutory systems or other settings with little scope for iteration.

A Test and Learn approach strengthens – rather than replaces – robust evaluation. This staged approach ensures that, by the time a full evaluation is commissioned, the intervention is sufficiently understood and mature to support credible conclusions about effectiveness and value for money.

Test and Learn is not a new way of working. It draws on practices and methods that are common to other disciplines – agile design,[2] systems thinking,[3] adaptive management, user research,[4] prototyping, experimentation and behavioural science. However, it is not widely used across government and current application tends to be as a one off rather than systematically embedded in programme design and delivery. Growing the use of Test and Learn activity across government offers untapped potential to ensure that public services meet the needs of the public, achieve the outcomes policymakers intend, offer value for money and lower the risk of trying new things.

The content in this guide has been informed by wider literature, policy documents and experts from across Whitehall experienced in Test and Learn delivery.

2 Agile is a delivery framework that structures work into short, repeatable cycles where teams design, build and release improvements.

3 An approach to understanding and addressing complex issues by looking at how different parts of a system interact and influence each other over time. Civil servants are increasingly using this approach in the various phases of the policy cycle.

4 Government Digital Service User Research is used in agile digital delivery and focuses on small, frequent rounds of research in every iteration, whereas Government Social Research covers a wide range of social research methods across policy design and evaluation.

HM Treasury

Chapter Two

What does Test and Learn mean?

2

Test and Learn is a way of working that centres around the use of feedback loops to check whether assumptions – that are key to policy success – are met in practice. Feedback can be generated through different methods and approaches such as participatory research, data analysis, prototyping and experimentation. Iterative feedback loops provide valuable information as to how policy interventions should be adapted to achieve the intended outcome.

Test and Learn activity is intended to home in on policy or service elements to assess how they are working in practice. The purpose is to develop and iterate policies and services to give them the best chance of success. It is not a substitute for evaluation, but a complementary approach that strengthens design, builds adaptive capacity and helps ensure that policies and services are fit for purpose before they are rolled out at scale.

7 key features of Test and Learn

1. **Delivered by multidisciplinary teams.** Working groups of 'do-ers' from central and local government, with expertise in policy, delivery, design, analysis and other disciplines such as system designers and behavioural insights should be convened to deliver Test and Learn activity. These teams work in partnership with appropriate stakeholders – including frontline staff, service users, delivery partners and others with relevant experience – to ensure that testing and learning are grounded in real-world conditions. This participatory approach strengthens problem understanding, design quality and buy-in.
2. **Starts with shared, measurable outcome/s**. Partners should agree and unify around a common outcome, which Test and Learn activity is then designed to support. This should be measurable and clearly stated at the offset. It should also relate to a tangible difference that the intervention is intended to make to people's lives.
3. **Based on an evidence-based understanding of the challenge.** Understanding of the challenge is built from multiple sources of evidence, including existing research, data analysis and insights generated through ongoing data collection. Test and Learn values diverse perspectives and local knowledge, while placing them within a wider system view. This helps ensure that interventions focus on the most important drivers of the problem, while remaining open to learning and adaptation as understanding develops.
4. **Tests assumptions where there are greater uncertainty and risk.** Issues that are critical to achieving key outcomes but have weaker evidence or greater uncertainty should be prioritised for Test and Learn activities. These assumptions often relate to behaviour change, delivery feasibility, unintended consequences, or value for money. Focusing on risk in this way helps teams reduce uncertainty early, avoid costly failure later and target analytical effort where it adds most value.
5. **Iterative.** Learning is generated through feedback loops that combine qualitative and quantitative evidence, often using multiple methods in parallel. Analysis is intensive and formative, meaning findings are reviewed as evidence emerges and used immediately to refine design, adjust delivery or revisit assumptions. This contrasts with one-off or retrospective analyses and supports continuous improvement.
6. **Supports both scaling and local adaptation.** Test and Learn generates insight that can inform the scale-up of interventions that work, as well as the tailoring of solutions to meet the needs of different places or contexts. These are distinct considerations: scaling tests whether an approach can work consistently across contexts, while local adaptation ensures interventions fit specific local conditions. Learning remains iterative in both pathways.
7. **Transparent and open.** Make processes, decisions and progress visible to stakeholders and the public as they happen, fostering transparency, collaboration and shared ownership.

Test and Learn is used as a shorthand throughout this document to summarise the above practices.

HM Treasury

Chapter Three

Why does Test and Learn matter?

3

It is impossible to predict with certainty how the public will interact with a policy, programme or service once it meets the realities of delivery, or what challenges will arise along the way. Assumptions made in the design stage are often only tested years later, by which point it can be extremely costly to discover they were wrong. For interventions where adaption is possible, gathering early feedback and insight helps to test assumptions and build evidence-based solutions that work in practice. This is true at every stage of policy development – whether you are facing a new challenge or trying to make a longstanding programme more cost-effective.

However, it is also important to acknowledge that Test and Learn is most useful where early signs of change can be observed quickly. When outcomes take months or years to emerge, Test and Learn can still be used, but the focus should be on testing early mechanisms or behavioural signals rather than long-term outcomes.

It is equally important to recognise that some types of work – such as infrastructure projects – require a fully planned and agreed design before delivery can begin. In these cases, a waterfall approach is both necessary and appropriate, because iteration during delivery is not feasible.

Where adaption is possible, the aim is to improve specific parts of the intervention as early as possible, rather than attempting to prove the effectiveness of the whole policy at the end of the policy cycle. Once an intervention has been refined and stabilised through iterative testing, it can then be assessed using robust evaluation approaches, and – if it proves effective – scaled up or developed through a place-based approach with greater confidence.

A Test and Learn approach is more suitable when:

- **The problem is complex or uncertain**, with outcomes shaped by multiple interacting factors (e.g. homelessness, reoffending, service engagement).
- **There are several plausible ways to intervene**, and evidence is needed to compare options before committing to scale.
- **Key assumptions about how change will happen are uncertain or weakly evidenced**, particularly across the outcome journey.
- **Delivery contexts or populations vary across places**, requiring learning about what works where and for whom.
- **Behaviour change mechanisms need to be tested in real-world conditions**, rather than assumed from theory or evidence drawn from other settings.

Test and Learn supports innovation by encouraging exploration and experimentation of multiple options, helping teams identify which are workable and most likely to improve outcomes. It also builds collaboration and buy-in from frontline staff, beneficiaries and delivery partners, ensuring interventions are grounded in real-world experience.

Through rapid feedback loops and continuous refinement, Test and Learn enables teams to start small and increase investment as confidence grows. By the time a policy is evaluated in full, it may have already gone through several cycles of iteration – giving it a far stronger chance of demonstrating positive impact.

In complex environments, where public needs, delivery conditions and policy goals continually evolve, testing assumptions early, learning quickly and adapting often is not just preferable – it is essential.

HM Treasury

Chapter Four

How Test and Learn supports robust evaluation

4

Test and Learn is intended to compliment established evaluation approaches, not to replace them. Its purpose is to strengthen policy design and delivery under uncertain conditions, while preserving the conditions needed for robust evaluation at later stages. Because Test and Learn involves iteration and adaptation, it is important to be clear how rigour, bias and causal inference are managed in practice.

Preventing bias as interventions are adapted

When an intervention is adapted during implementation, there is a risk that teams lose sight of which version is generating results, or that early positive findings are given more weight than mixed or negative evidence. Test and Learn manages this risk by being explicit upfront requiring teams to set out a clear Theory of Change, identify prioritised assumptions, and agree decision criteria before testing begins. Any changes made to the intervention are formally documented and used to refine the Theory of Change, rather than being absorbed informally. This approach ensures that early findings are interpreted with caution rather than treated as proof of success.

Maintaining rigour through proportional methods

Rigour in Test and Learn is achieved through proportionality, rather than applying the same evaluation method at every stage. Early testing focuses on understanding feasibility, delivery and mechanisms of change, using methods that are appropriate to the questions being asked. More formal evaluation designs are reserved for later stages, once interventions are sufficiently stable and well defined. This approach reduces the risk of drawing strong conclusions from weak or immature evidence, while still enabling timely learning to inform decisions.

Supporting causal evaluation over time

Test and Learn does not remove the need for causal inference; it helps create the conditions in which it becomes feasible and meaningful. By involving analysts early and considering evaluability during design and testing, teams can make implementation and rollout choices that preserve opportunities for counterfactual evaluation. Teams are expected to use appropriate designs – such as well-powered randomised controlled trials or credible quasi-experimental approaches – to estimate impact and value for money where feasible.

Why this matters

Without this sequencing, there is a risk that interventions are rolled out before they are sufficiently specified or evaluated too early to generate reliable findings. Test and Learn reduces this risk by clearly separating exploratory testing from summative evaluation, while ensuring that early learning strengthens, rather than undermines, later causal assessment.

In this way, Test and Learn functions as a disciplined approach to managing uncertainty, supporting robust evaluation by clarifying what is being tested, when, and how evidence should be interpreted.

The *Building Choices* case study below illustrates this in practice. Early testing focused on delivery feasibility, screening tools and behavioural mechanisms, resulting in a stabler, clearly specified intervention and a well-defined evaluation strategy, including a randomised controlled trial. In this way, Test and Learn functions as a way of working that protects rigour over time, rather than a shortcut around formal evaluation.

Case study

Testing uncertainties in the Building Choices programme

The Building Choices programme is His Majesty's Prison and Probation Service's (HMPPS) new Accredited Programme (AcP). AcPs are structured cognitive-behavioural interventions that aim to reduce reoffending by helping participants develop skills to manage the factors linked to offending behaviour. Building Choices was designed to offer a more person-centred, strengths-based and skills-focused approach, addressing a broad range of criminogenic needs. These are dynamic risk factors such as characteristics, behaviours or circumstances that are linked to criminal behaviour. If these factors improve, the person's risk of reoffending usually declines.

The programme was initially tested across three prison sites and two probation regions. This small-scale test was used to examine whether the programme could be delivered as intended, whether it reached the right target group, and whether it showed early promise in achieving perceived changes in outcomes, as outlined in the Theory of Change.

In line with the **Test phase (see section 6)**, the team focused on testing areas of greatest uncertainty and understanding the mechanisms of change within their delivery contexts. For example:

Testing uncertainties: The evaluation explored whether the new *Programme Needs Identifier (PNI)* correctly screened participants, whether the programme's group size and frequency of sessions were workable, and whether the preparatory 'Getting Ready' module was effective. These were all assumptions critical to the programme's success. **Methods used:** Qualitative interviews (multiple interviews with some participants) with prison staff, participants, court assessors and court staff to capture experience over time and to assess the Programme Needs Identifier (PNI) screening tool.

Examining mechanisms: The evaluation investigated how core elements of the programme – such as the *Success Wheel Measure (SWM)*, the Great Eight skills toolkit, and the use of skills logs – helped support changes in behaviour, such as improved emotional insight and problem-solving. **Methods used:** Participant-reported measures and outcome indicators to track early changes, e.g. in 'distance travelled' in skills domains like 'Positive Relationships' and 'Healthy Thinking'.

Considering context factors: Findings highlighted that delivery conditions influenced outcomes. For example, attrition was higher in community settings than in prisons due to participants facing greater external responsibilities and barriers. Staff also reported that the length of the Delivery Guide and challenges in delivering a new programme at pace shaped how well the programme could be implemented. **Methods used:** Examined variation in monitoring data, e.g. attrition (drop-out) by setting (e.g. comparing community vs prison settings), how delivery tools (e.g. delivery guides, timing, resource constraints) impacted how well the programme could be implemented.

The next step for the Building Choices team is to move into the **Grow phase (see section 6)**. With early uncertainties tested, mechanisms refined, and delivery adjusted, the programme is now in a more stable form that can be robustly evaluated at a larger scale. The evaluation plan includes a randomised controlled trial is underway, in tandem with a process evaluation, in early adopter sites to understand the impact of AcPs on short-term outcomes (such as improved impulse control, emotional regulation, problem solving and motivation).

As the Building Choices programme is expanded, researchers will evaluate whether the programme achieves its ultimate outcome of reducing reoffending. By structuring learning in this way, the Building Choices programme will build a strong and credible evidence base and guide decisions on expansion, tailoring for specific groups and adapting content. The Next Generation of Accredited Offending Behaviour Programmes: Building Choices.

HM Treasury

Chapter Five

How to prioritise where Test and Learn should be used

5

To apply a Test and Learn approach effectively, partners must first agree and clearly articulate the outcome a policy intervention is intended to achieve. The outcome/s should reflect the tangible differences that an intervention makes to people's lives and be measurable in real-world terms. The policy intervention and expected outcome/s should be articulated in a Theory of Change (see Magenta Book Chapter 2) that maps the pathway to impact and surfaces the key assumptions on which success depends. In line with the Nesta Test and Learn Playbook,[5] key assumptions might relate to:

- **Beneficiary relevance:** The intervention aligns with the realities, motivations and behaviours of the people it is intended to benefit.
- **Expected effect:** The intervention will produce the anticipated scale and nature of change if implemented as intended.
- **Economic viability:** The intervention is financially sustainable within current and projected resource constraints.
- **Delivery feasibility:** The intervention can be implemented effectively using available capabilities, systems and infrastructure.
- **Potential to scale:** The approach can be successfully expanded or adapted to reach a broader population or applied in different settings.
- **Unintended consequences:** The intervention will not create significant negative or unintended effects, including for groups not initially targeted. In practice, some unintended effects may emerge gradually, so ongoing monitoring and review is important.

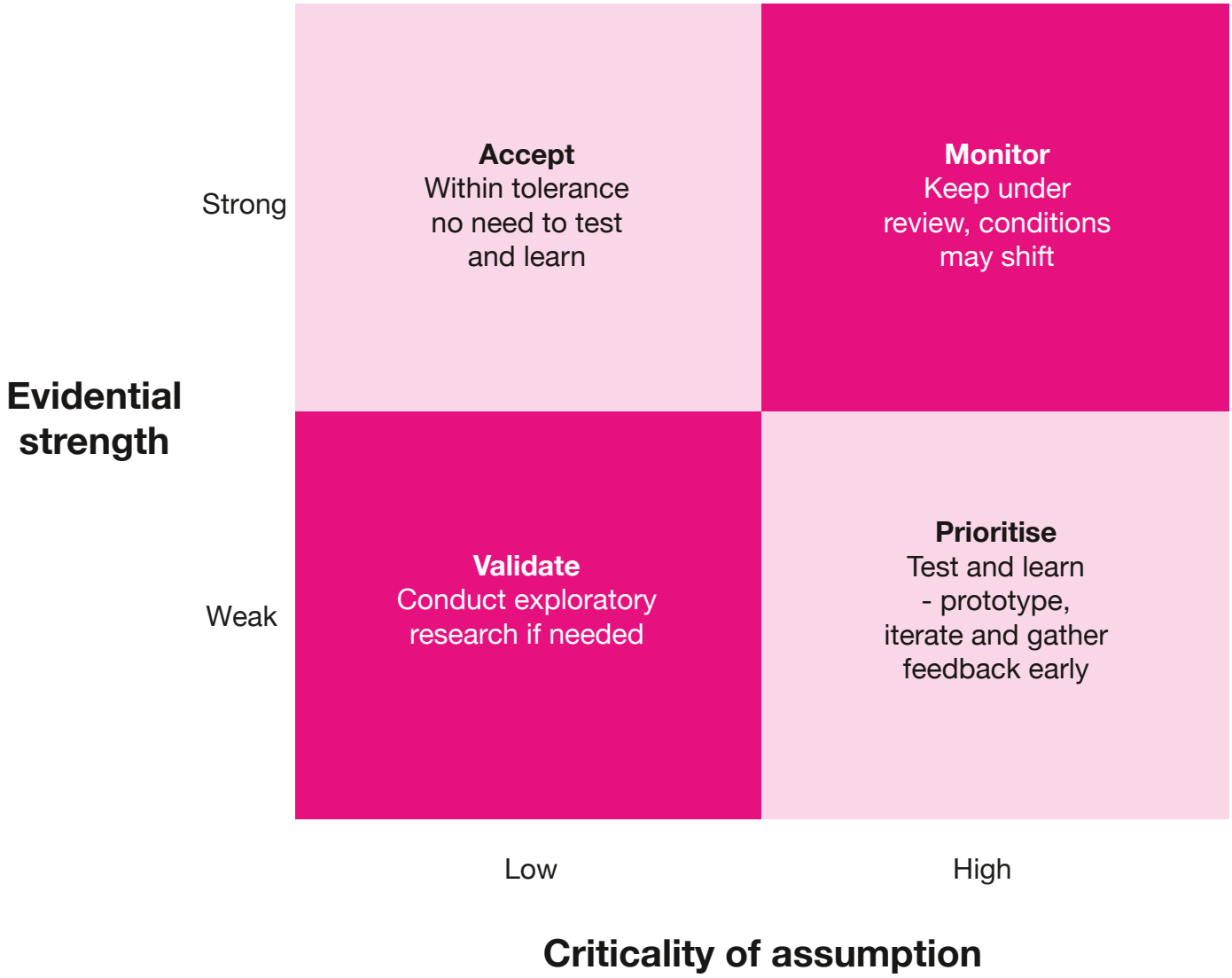
When there is significant uncertainty about how an intervention will work in practice – for example when user needs, delivery conditions or behavioural responses are not yet well understood – prototyping, iteration and rapid feedback allows teams to test critical assumptions early and refine design decisions before major investment.

To prioritise which components of an intervention could benefit from Test and Learn activity, assumptions can be categorised by two dimensions (see Figure 5.1):

- Their **criticality:** how much they matter to the intended outcome, and.
- Their **evidential strength:** how well-supported they are on a range of readily accessible past policy and other research, evidence-based literature and a knowledge of limitations/gaps in administrative or other data.

5 Nesta (n.d.) *Test and Learn: A Playbook for Mission-Driven Government*. Available at: https://www.nesta.org.uk/report/test-and-learn-a-playbook-for-mission-driven-government/ (Accessed: April 2026).

Figure 5.1: Prioritisation matrix



Test and Learn should not attempt to test every assumption equally. It is most effective when focussed on critical assumptions with a weak evidence base, for example, where there is high uncertainty, where a problem is still emerging or where existing evidence comes from a different setting and may not transfer in practice. In these situations, Test and Learn uses social research methods to clarify the problem, refine or adapt interventions and optimise solutions before undertaking robust impact evaluation of the resulting approach.

HM Treasury

Chapter Six

How to implement Test and Learn

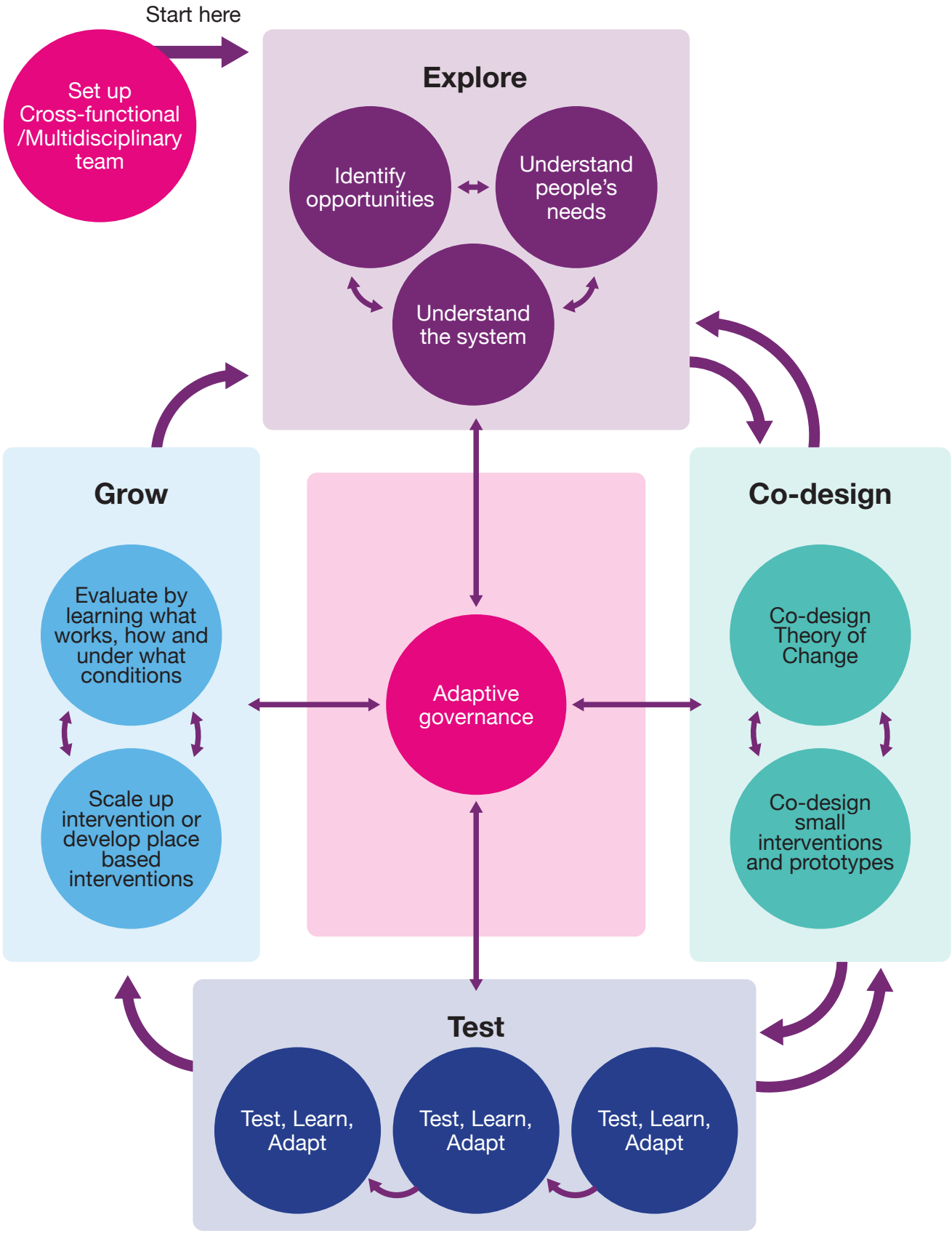
6

The Test and Learn framework (see Figure 6.1) shows how activity is organised around four connected stages – **Explore, Co-design, Test and Grow** – supported throughout by adaptive governance and multidisciplinary teams. The framework was developed through a review of existing literature and policy guidance, alongside iterative refinement with practitioners across government drawing on practical delivery experience.[6] Rather than a linear sequence, the diagram illustrates how learning flows between stages: insights from testing can prompt further Co-design, evaluation can inform decisions about scaling or local adaptation, and teams may return to exploration as understanding deepens. The purpose of this framework is to make explicit how evidence, feedback and decision-making are brought together over time. Each stage is described in more detail in the sections that follow, alongside practical guidance on methods, governance and decision points.

6 The framework draws on established literature on adaptive policy design, systems change, developmental evaluation, prototyping and internal cross-government practice:

- Lowe, T., Padmanabhan, C., McCart, D., McNeill, K., Brogan, A. and Smith, M. (2022) *Human Learning Systems: A Practical Guide for the Curious*. Centre for Public Impact.
- Ministry of Justice (2024) *Prototyping and Evaluating Policy Design: A Proportionate Framework*. Evaluation and Prototyping Hub, Analysis Directorate.
- Nesta (2023) *Test and Learn: A Playbook for Mission-Driven Government*. London: Nesta.
- Patton, M.Q. (2010) *Developmental Evaluation: Applying Complexity Concepts to Enhance Innovation and Use*. New York: Guilford Press.

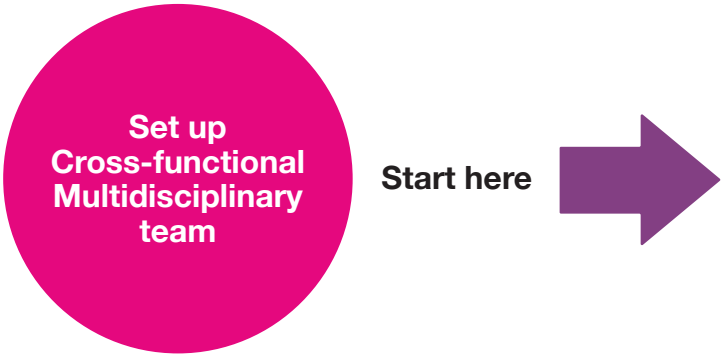
Figure 6.1: Test and Learn framework



6.1. Set up a cross-functional multidisciplinary team

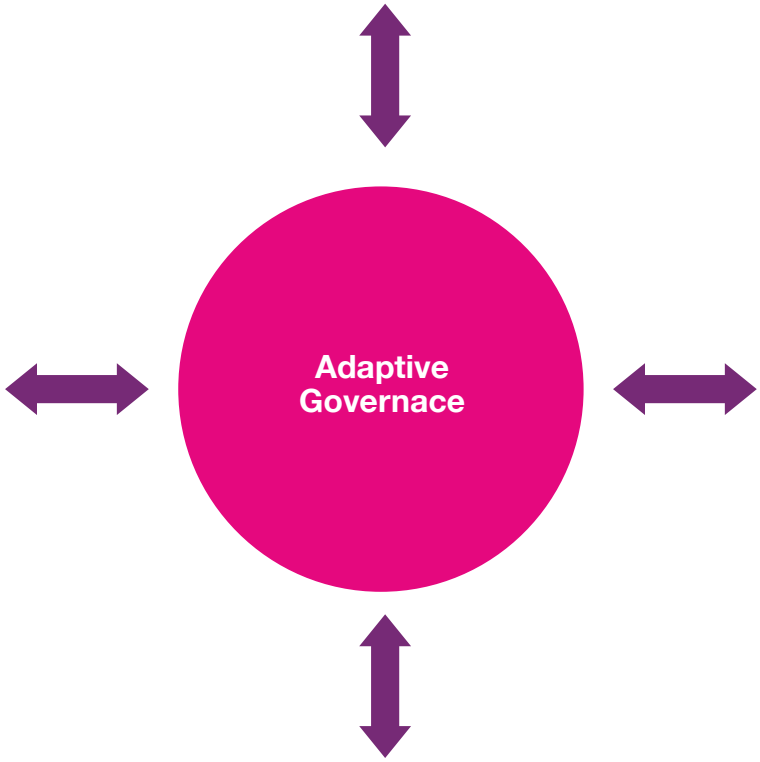
Collaboration is most effective when expertise is brought together from the start, rather than working in separate silos. Establishing a multidisciplinary team early in the process – drawing on knowledge from across government departments, local authorities, frontline services and a range of methodological expertise, possibly with external experts e.g. charities, academia and business – increases the likelihood of developing solutions that are practical, evidence-based and responsive to real needs. These teams should have the authority and permission to design solutions tailored to their local context, ensuring they reflect both the realities of delivery and the diversity of user needs.

Figure 6.2 The Set Up Stage



6.2. Adapting governance for Test, Learn, Grow

Figure 6.3: The Adaptive Governance Stage



To enable a Test and Learn approach, governance arrangements may need to be adapted to support iterative development, proportionate oversight and timely decision-making. The Government Principles for Agile Service Delivery[7] emphasise the importance of setting clear outcomes, empowering teams to act within defined parameters, and ensuring governance processes are lightweight, transparent and responsive to change.

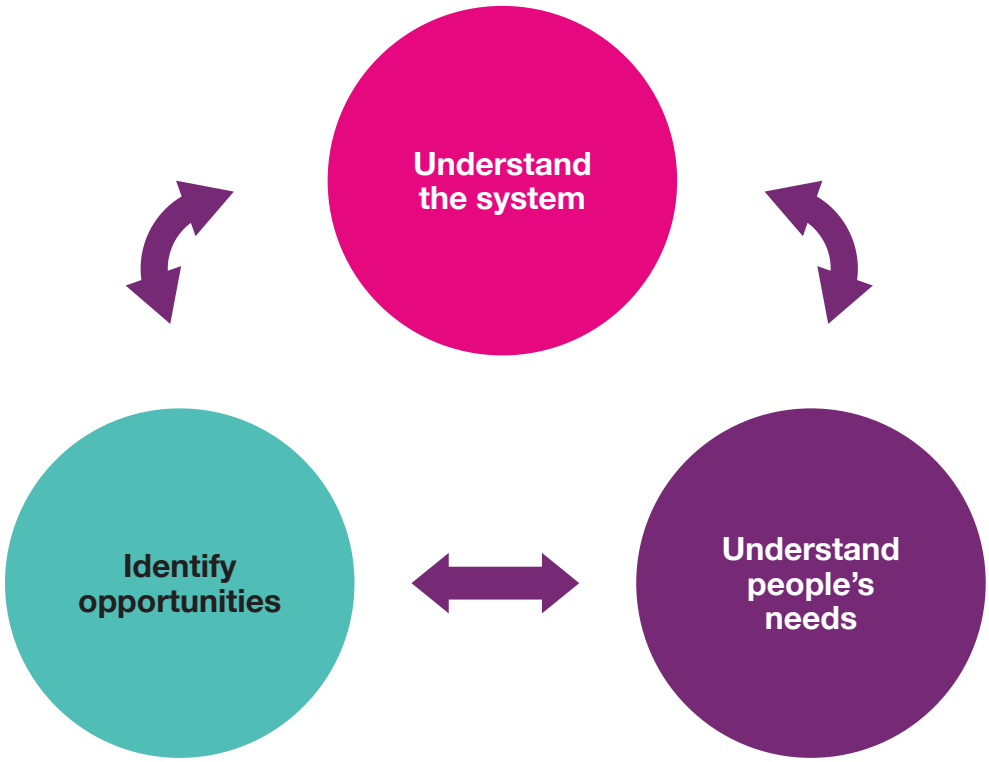
In practice, this means creating space for experimentation, accepting that not all tests will succeed, and basing decisions on evidence generated through the cycle rather than fixed plans made at the outset. Wider system barriers – such as rigid funding models, siloed decision-making, lengthy approval processes, challenges with data sharing and risk-averse organisational cultures – can hinder agility and delay the implementation of learning.

Addressing these constraints may require early engagement with senior leaders, commissioners and assurance bodies to agree proportionate controls, devolve decision-making authority, and embed a shared understanding of the value of iterative testing.

7 Gov.uk (n.d.) *UK Government Core Principles of Agile Delivery*. Available at: https://www.gov.uk/service-manual/agile-delivery/core-principles-agile (Accessed: April 2026).

6.3. EXPLORE: Building a shared system understanding

Figure 6.4: The Explore Stage



The Explore stage is the starting point for any Test and Learn initiative. It begins with a short, clear challenge statement – a description of the issue being tackled. A good challenge statement is built from evidence, data and insights gathered from those who use, deliver and are affected by a service or policy. Its purpose is to create a shared understanding of the problem. It should be specific enough to guide action, but broad enough to allow space for different solutions.

Next, the focus is on developing a shared understanding of the system in which the policy or programme operates. Rather than jumping straight into solutions, this stage focuses on identifying the system's actors, relationships, behaviours and contextual factors that shape the defined challenge.

This stage is collaborative, and it begins by engaging those closest to the issue – policy teams, delivery partners, service beneficiaries, academics – to surface the different perspectives and co-create a shared understanding of the system. Importantly, the issues raised through this engagement should be contextualised using wider evidence and data analysis, so that individual experiences are understood within broader patterns and system dynamics.

A range of social research methods can support this process (see Table 6.1). These include evidence reviews to summarise what is already known and to identify gaps and participatory approaches that actively involve stakeholders in understanding, designing and testing potential solutions. Other methods such as systems mapping, service blueprinting, soft systems methodology, actor mapping and behavioural diagnosis can help visualise how different factors interact and make visible the complexity of the challenge. Data analysis (i.e. reviewing administrative datasets, performance metrics, or linked data sources) can identify patterns, highlight variation across groups or places, and test whether locally raised issues are reflected more widely. Finally, developing personas – evidence-based profiles that represent different groups of people affected by the issue – and mapping their journeys helps ensure the team's understanding of the challenge is grounded in people's real-world experiences and needs.

Table 6.1: Example methods for the Explore phase

Category	Method	Purpose in Explore phase	Further information
System approaches	System mapping/causal loop diagrams	To visualise how actors, processes, and feedback loops interact; identify leverage points and unintended effects	Handling Complexity Guidance[8] and Systems Thinking for Civil Servants[9]
	Soft systems methodology (SSM)	To build a shared understanding of complex problems from multiple perspectives; define and refine the system boundary	Handling Complexity Guidance and Systems Thinking for Civil Servants
	Actor mapping	To identify key stakeholders, their relationships, roles and interests to understand power dynamics, incentives and potential blockers or enablers	Handling Complexity Guidance and Systems Thinking for Civil Servants
Participatory and qualitative approaches	Stakeholder engagement/ participatory workshops	To capture diverse viewpoints, surface assumptions and co-define the problem space	Handling Complexity Guidance
	Interviews and focus groups	To understand the context, motivations and barriers of those who deliver or experience a service	Magenta Book[10]
	Observation	To explore behaviours and contexts in situ, revealing how services operate in reality versus design	Magenta Book

8 HM Treasury (2025) *Supplementary Guide: Handling Complexity in Policy Evaluation*. Available at: https://www.gov.uk/government/publications/the-magenta-book/supplementary-guide-handling-complexity-in-policy-evaluation-html (Accessed: April 2026).
9 Government Office for Science (2023) *An Introductory Systems Thinking Toolkit for Civil Servants.* Available at: https://www.gov.uk/government/publications/systems-thinking-for-civil-servants/toolkit (Accessed: April 2026).
10 HM Treasury (2022) *The Magenta Book: Central Government Guidance on Evaluation*. Available at: https://www.gov.uk/government/publications/the-magenta-book (Accessed: April 2026).

Category	Method	Purpose in Explore phase	Further information
Behavioural diagnostic approaches	Behavioural diagnosis	To identify behavioural drivers, barriers and biases relevant to the policy issue; frame hypothesis for testing	Achieving behaviour change – a guide for national government[11]
Evidence synthesis	Rapid evidence reviews	To summarise research and identify where new evidence is needed	Magenta Book
Data and analytical approaches	Descriptive and exploratory analysis	To analyse administrative or performance data to identify patterns, variation and emerging issues	Magenta Book
	Social network analysis	A quantitative analytical method that uses data (e.g. interactions, referrals, partnerships, communications) to map and measure relationships between actors	Handling Complexity Guidance

11 Gov.uk (n.d.) *Behaviour Change: Guides for National and Local Government and Partners*. Available at: https://www.gov.uk/government/publications/behaviour-change-guide-for-local-government-and-partners (Accessed: April 2026).

Outputs from Explore might include:

- **A shared system view:** An agreed, high-level map showing the key elements, relationships and feedback loops within the system. This helps teams see how different parts connect, where constraints or dependencies exist, and where change might have the greatest effect.
- **An understanding of what shapes behaviour:** An assessment of the factors that influence what people do – including individual characteristics (such as skills, confidence, and habits) and environmental factors (such as processes, resources, incentives, and social norms). This output provides a structured picture of behavioural drivers and barriers and is particularly valuable when the challenge involves changing how people act or make decisions.
- **A clear, collectively owned understanding of beneficiaries' needs:** A summary of who is affected by the issue, what matters to them, and how their needs differ across groups. Beneficiaries often include a wide diversity of people — and sometimes non-human actors within the wider system — so capturing a range of perspectives is essential. Tools such as CATWOE[12] from Soft Systems Methodology can help surface these differing worldviews.
- **A clear statement of what is not yet known:** Identifying gaps in evidence and uncertainties about the problem or delivery context, which will guide what needs to be explored, prototyped and tested in subsequent stages of the Test and Learn process.
- **An assessment of key opportunities and leverage points:** A concise analysis identifying areas within the system where well-designed interventions could create significant positive change. This helps prioritise where to focus testing and learning effort.

12 Checkland, P. and Poulter, J. (2010) 'Soft systems methodology', in *Systems Approaches to Managing Change: A Practical Guide*. London: Springer.

Case study

Systems-wide Evaluation of Homelessness and Rough Sleeping: Criminal Justice and Homelessness

This example aligns with the Explore stage because it brings together evidence from across criminal justice, housing and homelessness systems to build a shared understanding of how interactions, transitions and institutional arrangements shape homelessness outcomes, before any specific solutions are designed or tested.

This case study sits within the wider Systemswide Evaluation of Homelessness and Rough Sleeping, a programme of research and evaluation that seeks to understand how homelessness is shaped by interactions across multiple policy areas and the broader socioeconomic context. Rather than focusing on individual services or interventions in isolation, the evaluation examines how outcomes for people experiencing homelessness emerge from the way different systems operate together.

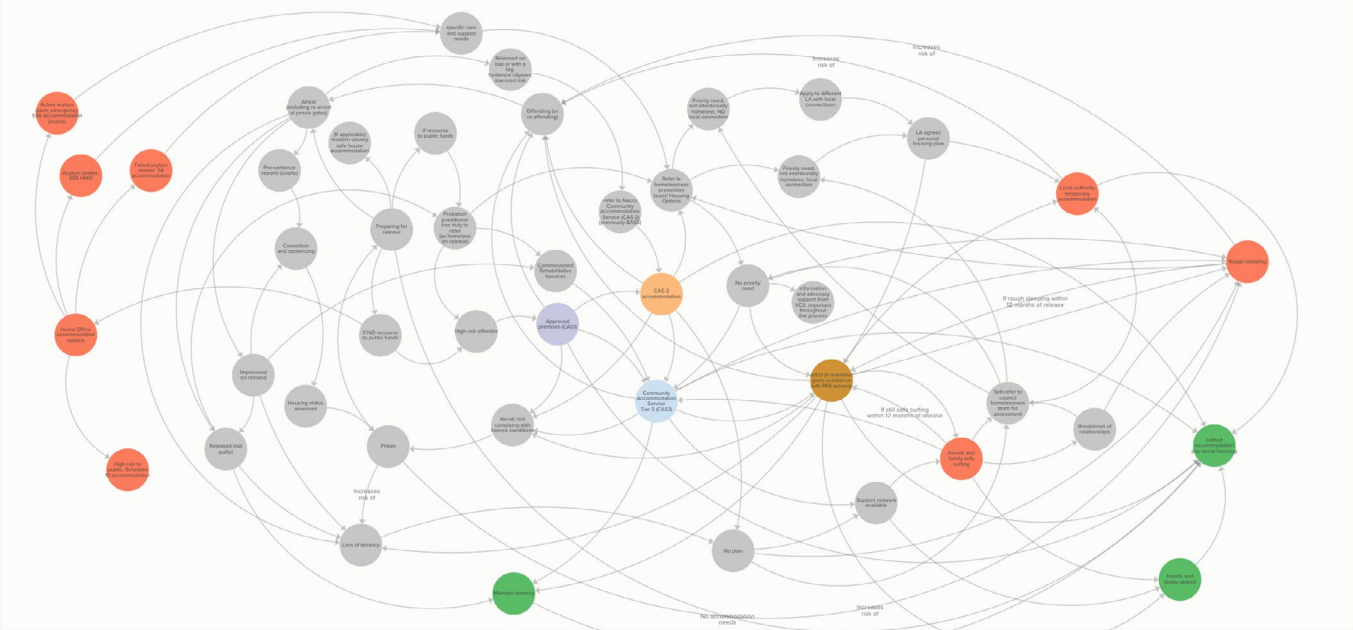
The evaluation has produced a range of reports exploring different parts of the homelessness system and its interactions with other policy areas, including asylum, housing and criminal justice. Findings from the evaluation have already informed policy developments set out in the National Plan to End Homelessness, including changes to funding models to strengthen prevention and a review of legislation and guidance affecting housing allocations. The final report, expected in 2027, will draw together evidence from across the evaluation and the Test and Learn programme to identify the most effective and impactful levers for sustainable systems change.

One critical area of interaction explored within the evaluation is the relationship between homelessness, rough sleeping and the criminal justice system. Homelessness and rough sleeping are persistent and complex social issues, with people experiencing homelessness disproportionately likely to encounter police, courts, prisons and probation services. Evidence consistently shows a strong relationship between homelessness, reoffending and repeated criminal justice involvement. Housing instability can increase the likelihood of contact with the criminal justice system, while involvement with that system can, in turn, disrupt access to stable accommodation and support. Understanding how these feedback loops operate across systems is essential to addressing homelessness and rough sleeping among this cohort effectively.

The criminal justice deep dive employed a systems-wide analysis that brought together evidence from across the criminal justice system, housing and homelessness services, and local government, recognising that outcomes for prison leavers are produced through the interaction of responsibilities, decisions and practices across these systems. The research combined multiple sources of evidence, including administrative data, qualitative research and practitioner insight, to examine how people move through police custody, courts, prisons and probation, and how these pathways intersect with housing and homelessness services. By following people's journeys across systems, the analysis sought to understand where risks are concentrated and how system design and decision-making shape outcomes.

Bringing evidence from different parts of the system together made visible the fragmentation between criminal justice, housing and homelessness services and the way responsibilities are distributed across national departments, local authorities and delivery organisations. The findings highlighted how differing priorities, funding arrangements and operational pressures shape decision-making within individual services, often with limited coordination across system boundaries. This system-wide perspective demonstrated how homelessness and reoffending can be reinforced by existing institutional arrangements, particularly at key transition points where risks are highest. This led to the identification of gaps in support for certain groups such as women, those on remand and those serving shorter sentences.

Figure 6.5: A systems map of the homeless and rough sleeping evaluation



The systems map is available at this link: Routes into accommodation or homelessness for adults leaving prison

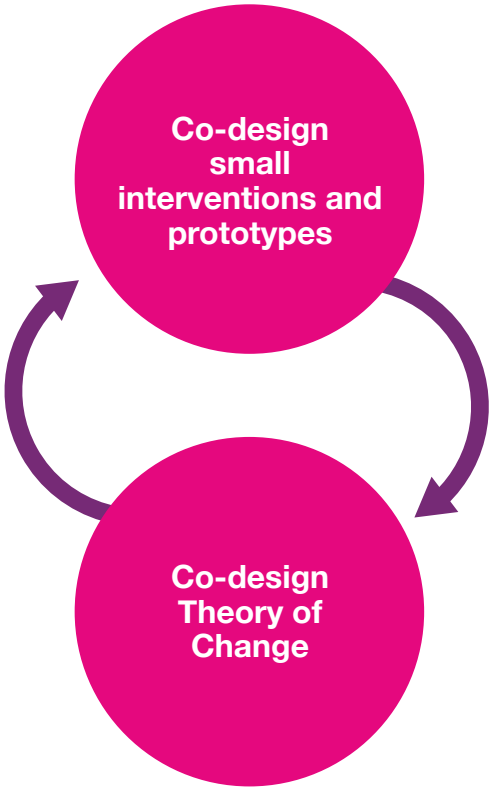
Checklist: Practical steps for the Explore stage

- ✓ **Engage key stakeholders:** involve policy teams, delivery partners, service beneficiaries, academics and others with insight into the issue.
- ✓ **Develop a challenge statement:** use evidence, data and insights gathered from those who use, deliver and are affected by a service or policy to create a shared understanding of the problem.
- ✓ **Map the system:** identify key actors, current actions and experiences, relationships, behaviours, feedback loops and contextual factors influencing the challenges by using structured methods such as rich pictures, systems mapping, service blueprints, journey maps, personas, soft systems methodology or actor mapping.
- ✓ **Understand what shapes people's behaviours:** identify individual influences (e.g. skills, confidence, habits) and wider system factors (e.g. the processes, infrastructure, social norms) that shape actions.
- ✓ **Understand beneficiaries' needs:** capture the diversity of needs, recognising that beneficiaries may include a wide range of people and, in some cases, non-humans.
- ✓ **Identify opportunities, constraints and leverage points:** highlight areas where well-designed interventions could have a disproportionate impact.
- ✓ **Build a shared understanding:** ensure findings are collectively owned by all involved, creating a strong foundation for the next phase.

6.4. CO-DESIGN: Including those who use or deliver a service

The Co-design stage is a collaborative approach to developing policy and services that involves the people who deliver and use them – including frontline staff, members of the public and partner organisations – directly in the design process. It recognises that those closest to a service or policy often hold unique insights into its challenges and opportunities. By working together from the earliest stages, teams can generate solutions that are practical, acceptable and effective, grounded in real-world needs and constraints.

Figure 6.6: The Co-design Stage



The first step is to generate multiple ideas that address the key problems and opportunities identified in the Explore stage. Existing evidence – such as performance data, evaluation findings, academic research or best practice from other programmes – provides a foundation for understanding what has worked elsewhere, common pitfalls and measurable success criteria. It can also be valuable to look to other sectors or situations facing similar challenges, identifying concepts that can be adapted for the current context. This helps the team to set out the initial logic for why intervening at one point in the system might be more effective than another. The logic is usually set out in an initial Theory of Change which will draw on the Explore assessment and Co-design ideas. Many What Works Centres and the Evaluation Registry host evidence repositories and practice guides that can be valuable sources when programmes begin to explore the available evidence base.

In practice, it is important to recognise that Co-design does not mean narrowing immediately to a single solution. Complex issues typically involve multiple, interrelated aspects and effective solutions often require exploring several options in parallel. Teams should therefore expect to design and test a range of prototypes – sometimes targeting different dimensions of the same issue, sometimes addressing different but connected challenges within the wider system. At any given point, a team may be progressing several prototypes simultaneously. This approach acknowledges the complexity of real-world delivery and increases the likelihood of arriving at interventions that are both practical and impactful.

Initial ideas can be turned into very basic ‘paper’ prototypes, such as concept scenarios and storyboards. The aim is to make the idea tangible enough for people to react to, so teams can gather rapid feedback at an early stage. Engagement with frontline staff, service users and partner organisations ensures that prototypes are enriched with lived experience, revealing assumptions, practical barriers and opportunities not visible in the data alone. Based on this feedback, teams can narrow down their concepts to a smaller number of prototypes to develop further.

> Prototypes are intentionally quick and low cost, enabling teams to explore how an idea might work without significant investment. Maintaining an open, iterative mindset is essential, treating each prototype as an opportunity to learn rather than as a final product.

The next step is to test selected prototypes in a real-world setting, in a controlled and time-limited way. This could involve trialling a version in an ‘offline’ environment that mirrors operational conditions or running it at a very small scale within a live service. The purpose is not to generate definitive evidence, but to see how people interact with the prototype and to surface practical issues early. At this stage, **qualitative and mixed-methods approaches** are often most useful.

Techniques such as **observations, shadowing** and **short interviews or informal conversations** with users and staff help capture immediate reactions and uncover points of confusion, workarounds or unexpected barriers. **Quick surveys** can provide structured feedback on satisfaction, usability and perceived impact. **Data analysis** can help teams to spot patterns and identify which design choices are working or where adjustments are needed. Combining these approaches enables teams to understand not just *what* is happening, but *why*.

Examples of methods for prototype testing at the Co-design stage

When testing prototypes at a very small scale, the goal is to **see how they work in practice and gather rapid feedback**. Given the stage of testing, a mix of light-touch qualitative and quantitative methods is likely to be proportionate. Because prototypes are designed to expose obvious issues – such as points of confusion, barriers to use or unexpected behaviours – findings often emerge quickly without the need for large samples or complex analysis. The emphasis is on generating enough insight to inform the next iteration, rather than producing definitive or generalisable evidence.

- **Observation and shadowing:** Watch how users or staff interact with the prototype in real time to identify confusion, shortcuts or unintended behaviours.
- **Data analysis:** Include simple data analysis, such as tracking take-up, engagement levels or drop-off points to spot patterns quickly.
- **Think-aloud testing:** Ask participants to narrate their thoughts while using the prototype, revealing hidden assumptions and usability issues.
- **Short interviews or informal conversations:** Capture immediate reactions and insights from staff, service users or delivery partners after trying the prototype.
- **Focus groups or group feedback sessions:** Bring together users or frontline staff to discuss what worked, what didn't, and how the prototype could be improved.
- **Quick surveys or feedback forms:** Gather structured data on satisfaction, ease of use, or perceived impact; useful for spotting patterns across participants.
- **Journey mapping:** Work with participants to chart their experience step by step, highlighting pain points, enablers and emotional responses.
- **Service walkthroughs/role-play:** Rehearse how the service would operate end to end, with users or staff playing roles to expose gaps and opportunities.

Top tip: Use a combination of methods where possible – for example, pairing short surveys with observations – to build a fuller picture of how the prototype is performing and why.

These methods can be used to assess prototypes against agreed success criteria such as user satisfaction, operational feasibility, anticipated costs and alignment with policy objectives. Prototypes which show clear potential can make rapid, evidence-based adjustments before investing further and retesting; those that do not should be stopped, with learning documented for future reference. This stage of the Test and Learn process therefore focuses on both adapting interventions and optimising solutions.

If early testing shows that ideas are not leading to the intended behaviours or outcomes, this should trigger a deliberate decision to return to the Explore stage. Revisiting the system view and considering alternative explanation and opportunities at this point ensures effort is redirected towards solutions with a stronger likelihood of success.

The final step is to revisit and refine the initial Theory of Change, considering what has been learned. The updated Theory of Change draws on both existing evidence and the findings from prototype testing to provide a more confident account of how the proposed intervention is expected to lead to change. It sets out the causal pathway between activities (what is done) and desired outcomes (what is achieved), identifies the assumptions that remain critical and considers risks and mechanisms for change. This provides a clear and robust framework for developing testing plans on critical evidence needs in the Test phase.

Checklist: Practical steps for the Co-design stage

- ✓ **Engage key participants:** Involve frontline staff, service users, and partner organisations from the outset.
- ✓ **Generate ideas:** Use evidence from the Explore stage (data, evaluations, research, best practice) and inspiration from other sectors to create multiple potential solutions.
- ✓ **Create simple prototypes:** Develop low-cost, low-effort models (e.g. storyboards) to make ideas tangible.
- ✓ **Gather early feedback:** Test prototypes informally with stakeholders to uncover assumptions, barriers and opportunities.
- ✓ **Narrow down options:** Select the most promising prototypes based on feedback for further development.
- ✓ **Test in real-world settings:** Trial selected prototypes in a controlled, time-limited environment (offline or small-scale live test).
- ✓ **Observe and collect feedback:** Capture qualitative insights and quantitative data on usability, feasibility and impact.
- ✓ **Assess against success criteria:** Evaluate against agreed measures such as user satisfaction, operational feasibility, cost-effectiveness and policy alignment.
- ✓ **Refine or stop:** Improve and retest prototypes with potential – stop those that do not meet criteria, recording lessons learned and exploring explanations.
- ✓ **Develop and refine a Theory of Change:** Set out the causal pathway, assumptions, risks and mechanisms to inform the Test phase.

Case study

Co-designing a ‘working week’ prototype in prisons

Prisons often struggle to provide enough purposeful activity, leaving prisoners locked in their cells for long periods with limited opportunities for work, training or education. The Prisons Strategy White Paper (2021) highlights evidence suggesting that regular purposeful activity can improve wellbeing, reduce tension and violence, and lower the risk of reoffending.

The MoJ Offender Employment Policy Team aimed to extend the working week to 35 hours and make workshops cost-neutral by improving profitability. From the start, the team were aware of multiple interdependent challenges, including staffing, safety, external business support, sufficient external orders, workshop capacity and prisoner access to meals, legal visits and exercise. The MoJ Evaluation & Prototyping Team advised focusing first on the greatest uncertainty – whether a full working week could be operationalised and run safely – and testing it in a small, live environment before making any major investments. The project team worked with operational staff at the prison to design how this might work.

A one-week prototype was run in a single workshop at one prison, using existing facilities. Packed lunches were provided to enable prisoners to remain in the workshop all day, and operations were observed closely to assess staffing, safety and the impact on the wider regime. The results were immediate and valuable: the team said they learned more in one week than they would have in six months of planning. The test highlighted practical issues such as scheduling, supervision, mealtime logistics and safety protocols, all at minimal cost.

Early assumptions in the ‘working week’ prototype were that prisoners could remain in workshops throughout the day, with lunch provided on site, without disrupting the wider regime or access to essential services. Live testing showed that these operational elements were challenging in practice. Lunch was delivered earlier than planned, supervision arrangements were inconsistent, and the movement and escort capacity required to support the longer working day meant that some prisoners found it more difficult to attend essential appointments, such as healthcare or legal visits, without missing workshop time. Access to fresh air and other core regime activities was also reduced.

These pressures contributed to lower attendance and increased drop-out during the week. None of these effects were apparent during planning, but they became immediately visible through in-situ testing and conversations with instructional officers, kitchen staff, prisoners and operational teams. The findings demonstrated that lunch arrangements, movement, supervision and escort capacity were critical design constraints, rather than secondary operational considerations.

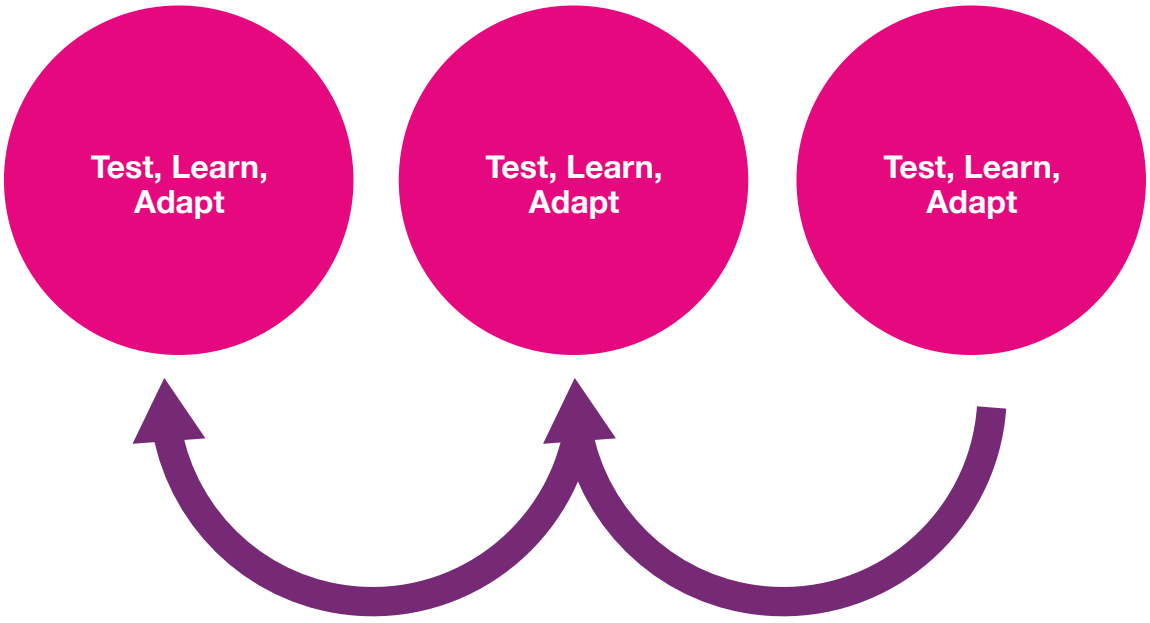
This led to a further three-month Co-design phase where the project team and staff on site worked to resolve the practical issues highlighted, prior to moving into the Test phase.

Co-design benefits

- Involved all stakeholders from the outset, building trust and buy-in.
- Turned ideas into tangible, testable changes in a live, time-bound setting.
- Provided quick real-world insight – it took four weeks from start to finish to run the first prototype.
- Exposed hidden barriers and assumptions, enabling rapid adjustments.
- Provided evidence to refine and retest the approach over a three-month period.
- Strengthened collaboration and momentum, leading to more resilient solutions.

6.5. TEST: Testing areas of greatest uncertainty

Figure 6.7: The Test Stage



The Test stage focuses on testing and refining the most critical or uncertain parts of your Theory of Change in a real-world setting. This phase recognises that you may not need – or be able – to test the entire policy or service in one go. Instead, the emphasis is on examining specific components, mechanisms or assumptions that are most risky, uncertain or vital to the success of the intervention. By isolating and testing these elements, teams can build robust evidence about what works, adapt their design accordingly, and avoid investing in large-scale implementation before key uncertainties are resolved. However, it's useful to also collect early evidence on whether the intended effects are starting to happen, to guide iteration and to decide whether more robust testing is justified later (in the Grow phase). In the Test stage, the focus is on whether the short-term outcomes or behavioural signals that underpin your Theory of Change are emerging.

Deciding what to Test will be informed by your Theory of Change from the Co-design phase. Consider questions such as:

- How are needs met for beneficiaries?
- What is the perceived value of the policy?
- What needs are not being met?
- What enabling and obstructing factors are emerging that are likely to be important?
- How has the context influenced delivery?
- Are the core mechanism(s) changing behaviour as anticipated?

Three key steps in the Test Stage

- **Start with your riskiest assumption(s):** These are those most critical to success but least supported by evidence (see Chapter 5). They are the potential 'showstoppers' identified in the Explore stage and the Co-design Theory of Change, that must be tested early to avoid costly failure later. Keep tests focused – avoid trying to test everything at once and instead use the right social research method to generate the specific insights you need to inform the next iteration. As confidence grows and the highest-risk assumptions are addressed, attention can shift to testing other, less critical uncertainties – recognising that new questions are likely to emerge as interventions evolve.
- **Select proportionate methods:** Where large investments or significant potential harms are at stake, rigorous testing may be justified even at an early stage (for example, using a randomised or quasi-experimental design to confirm a critical mechanism). Conversely, where the risks are lower, less rigorous social research methods may be sufficient.

- **Combine methods where possible:** No single method will capture the full picture of whether an intervention is working as intended. Combining quantitative approaches such as small-scale trial data, uptake data, A/B tests (see Table 6.2) with qualitative approaches such as interviews, observation, user diaries provides both breadth and depth of understanding. Mixed-method approaches also help uncover unintended consequences that purely quantitative or purely qualitative designs might miss. The aim is to generate richer, more reliable insights to guide decisions at each iteration.

  Social research methods used in the Test stage can take many forms (see Table 6.2). For example, if your intervention relies on voluntary participation, you might run small-scale trials to understand whether a sufficient proportion of the target audience engages, and under what conditions. This can quickly help you understand the financial viability of your intervention.

Mechanism experiments can be used to test why an intervention might work by isolating specific causal pathways. For example, whether a specific type of support session increases attendance at follow-up appointments, or a redesigned communication improves understanding of key information or personalised coaching rather than financial incentives is the element that drives sustained employment in a welfare programme.[13]

Testing mechanisms can build confidence in a policy before developing more expensive elements of the policy. These targeted tests help confirm whether the causal pathway assumptions in your Theory of Change are realistic and achievable.

13 Ludwig, J., Kling, J.R. and Mullainathan, S. (2011) 'Mechanism experiments and policy evaluations', *Journal of Economic Perspectives*, 25(3), pp. 17–38.

Table 6.2: Testing assumptions – methodological examples

If you want to test this assumption:	Examples of methods that could be used:[14]
Will people engage with the service when it is offered?	**Small-scale tests** in limited locations as an initial step with structured user feedback; analysis of **uptake rates** using administrative data; and **A/B testing** of invitations or access routes.
Are frontline staff able to deliver the service as intended?	**A series of qualitative sprints using process evaluation techniques** (focused on specific elements not the whole programme) e.g. fidelity checks, delivery logs; **ethnographic observation; interviews/focus groups** with delivery partners.
Does the intervention change user behaviour in the way we expect?	**Mechanism experiments: randomised controlled trials or quasi experimental methods** that isolate specific components such as type of support, timing, or communications. These tests provide early evidence on causal pathways but should be followed by more robust evaluation if the mechanism appears promising.
Are there unintended consequences (positive or negative)?	**Qualitative interviews, ethnographic research,** or **service mapping** to identify knock-on effects; monitoring / analysis of related administrative datasets to spot system-level shifts; **randomised control trials** can reveal unexpected changes in behaviour or outcomes.
Is the intervention feasible and sustainable in practice?	**Small-scale feasibility tests** in different contexts (urban/rural, large/small providers); **cost-tracking studies** (tracking actual resources used and compare with original budget assumptions); structured **feedback from delivery staff**. These early checks should inform whether the approach is ready for more robust evaluation.

14 Teams are not expected to use all methods listed. Methods should be selected proportionately, based on the assumptions being tested, the decision at hand, and the time and resources available. Using more methods is not inherently better and may increase costs without improving insight.

If you want to test this assumption:	Examples of methods that could be used:[14]
Are intended short-term outcomes or behavioural changes identified in the Theory of Change starting to appear?	Use **proxy or mechanism-level measures** (e.g. self-reported confidence, observed engagement, short pre/post surveys); analyse **administrative data trends** for early shifts; run **nimble randomised controlled trials or quasi-experiments** on specific components; use **Qualitative Impact Protocol (QUIP)** or **contribution stories** to explore perceived change pathways. These early signals should guide iteration and preparation for a more robust evaluation design.
Is there sufficient value added to justify scaling?	**Early cost-effectiveness analysis** using data from initial tests; comparison with **benchmarks** from similar interventions; **survey data** on user outcomes. These should inform the case for a full impact evaluation.

The Test phase is also critical for identifying unintended consequences or practices that may undermine the intended outcomes. Testing on a small scale allows these effects to surface early, before they become embedded in wider delivery. These might include increased demand on a related service, negative impacts on certain user groups, or operational workarounds that reduce effectiveness. Documenting these findings enables teams to make informed adjustments to the design, mitigate risks and strengthen the overall programme theory.

Deciding what to Test will be informed by your Theory of Change from the Co-design phase. Consider questions such as:

- How are needs met for beneficiaries?
- What is the perceived value of the policy?
- What needs are not being met?
- What enabling and obstructing factors are emerging that are likely to be important?
- How has the context influenced delivery?
- Are the core mechanism(s) changing behaviour as anticipated?

Examples of methods that can be used in the Test phase

- **User testing** involves observing and gathering feedback from people as they interact with a service, policy tool or prototype. It helps teams understand whether interventions are usable, acceptable and effective from the perspective of those who will experience them. Sessions are often short, focused and iterative – users are asked to complete tasks or respond to scenarios while researchers note where they succeed, struggle or react in unexpected ways. This direct insight into the lived experience of delivery can reveal barriers, assumptions and opportunities for improvement that would not emerge from quantitative data alone.
- **Qualitative sprints** are intensive, short bursts of research (often one to two weeks) designed to rapidly generate insight that feeds directly into adaptation. Techniques may include structured interviews, focus groups, journey mapping or ethnographic observation. The aim is not to produce a full evaluation report but to answer specific questions quickly, such as "Do frontline staff understand the new process?" "Are users clear about what is expected of them?" and "What unintended effects are emerging in practice?"
- **A/B testing** is a method typically used in digital services to compare two or more versions of a service feature to determine which performs better. However, the same principles can be applied to public policy. A/B tests will randomly assign a particular design/messaging strategy/feature to different groups of users. Response rates can then be compared to see what is most effective – for example testing two different mentoring approaches to see which leads to higher engagement. It is best suited to decisions about specific design features or delivery mechanisms, rather than evaluating whole programmes. As such, it can be a valuable tool in the Test phase of test and learn, where the aim is to understand which design choices show the most promise before scaling up. A/B testing is usually implemented in a similar way to a randomised controlled trial.
- **Online test experiments** involve running randomised tests of simplified versions of an intervention with an online participant pool or through digital experimentation tools. They allow teams to compare different design options or behavioural prompts quickly and at low cost, helping to identify which versions show the strongest potential. These tests provide early signals about which mechanisms may work and help narrow down a wider set of ideas before committing to more resource-intensive real-world testing. They should be used as an initial step in evidence building, not as a substitute for robust evaluation in operational settings.
- **Ecological momentary assessment (EMA)** is a method that collects feedback from participants in real time as they experience an intervention. Instead of relying on retrospective surveys or interviews, EMA prompts users – often through mobile devices or short diary entries – to record their thoughts, feelings or behaviours at multiple points during the day or week. This approach reduces recall bias and captures context-specific insights that might otherwise be missed, such as how people respond in different settings, times of day or emotional states. In the Test phase, EMA can help teams understand whether a prototype is being used as intended, how it fits into people's daily lives, and what moments create friction or enable success.

- **Nimble randomised control trials (RCTs)** can be used to test and improve the efficacy of specific feature of the policy's delivery to provide evidence on critical design choices. As in an RCT participants are assigned to intervention and control groups. However, whereas a full-scale effectiveness RCT tests the overall impact of the policy, nimble RCTs focus on improving the efficacy of the policy by testing features within it. This makes them smaller and faster to run. Nimble RCTs still require careful design to ensure they have appropriate statistical power
- **Efficacy RCTs** can allow teams to test whether the causal pathway is operating as intended and whether the intervention triggers the expected early changes under controlled conditions. This type of trial provides strong evidence on *whether* the intervention can work, even if results cannot yet be generalised to wider settings. Such designs are especially useful when longer-term outcomes will take years to emerge, when decision-makers need credible early evidence before committing further investment, or when isolating mechanisms is critical to refining the model ahead of larger-scale evaluation.
- **Bayesian adaptive trials (BATs)** support a '*learning as you go*' approach, making them well suited to agile policy design and iterative evaluation.[15] Like a traditional RCT, participants are randomly allocated to different options (or 'arms'), but the Bayesian approach updates the probability that each option is effective every time new data is observed. This means trial design features – such as sample size, allocation ratios, or whether an arm is dropped – can be adapted in real time to make the trial more efficient and ethical.
    - **Efficiency:** Bayesian trials can reach reliable conclusions with smaller sample sizes because they use all available information, including prior knowledge.
    - **Flexibility:** BATs allow adjustments as evidence emerges. This makes them better suited to uncertain or evolving policy contexts.
    - **Early stopping:** If an option clearly works (or clearly doesn't), the trial can stop sooner, reducing wasted time and resources.

Decisions about whether to continue, adapt or stop prototypes should be based on evidence collected during testing and assessed against agreed success criteria, defined in the Explore stage. Considering decision points in advance – including what evidence would be needed to justify continuation – helps guard against bias and over-commitment to an approach simply because time and effort have already been invested. This discipline ensures that decisions based on findings from the Test phase are transparent, consistent and grounded in evidence, even when they involve stopping a promising but unworkable approach.

Learning from the Test stage should be systematic and iterative. Evidence from these targeted tests should feed back into the Theory of Change, refining the description of how change is expected to happen and updating assumptions, risks and adapting the approach to meet success criteria. This is likely to occur at the end of the Test phase prior to decisions about whether to move into

15 Cripps, S., Lopatnikova, A., Afshar, H.M. et al. (2025) 'Bayesian adaptive trials for social policy', *Data & Policy*, 7, e27. doi: 10.1017/dap.2024.97.

the Grow phase or if the funding is phased, this might occur at the end of each funding cycle. This ensures that when the intervention moves towards larger-scale evaluation or full implementation, that is based on Test and Learn insights and has been shaped by real-world learning.

You should only move from the Test phase into the Grow phase once the intervention has reached a steady state – where its design, delivery and core processes are stable and operating as intended. At this point, the focus shifts from refining and improving the approach to assessing its overall effectiveness, impact and value for money.

Checklist: Practical steps for the Test stage

- ✓ **Identify critical uncertainties:** Select the most risky, uncertain or important parts of your programme theory to test.
- ✓ **Design targeted tests:** Plan small-scale trials or mechanism experiments to focus on specific components, assumptions or causal links.
- ✓ **Use real-world settings:** Test in operational or realistic environments that mimic the operational setting to capture authentic user and staff experiences.
- ✓ **Collect proportionate evidence systematically:** Gather both qualitative feedback and quantitative data to assess usability, feasibility and impact.
- ✓ **Monitor for unintended effects:** Identify negative impacts, operational workarounds or knock-on effects on other services early.
- ✓ **Set decision points and criteria in advance:** Agree in advance what evidence would justify continuing, adapting or stopping the prototype. Using predefined criteria and transparent decision-making reduces bias and avoids over-commitment to unworkable ideas.
- ✓ **Adapt or stop based on evidence:** Refine and retest prototypes that show potential; discontinue those that do not meet success criteria.
- ✓ **Feed learning into the Theory of Change:** Update assumptions, risks and causal pathways to reflect real-world insights.

Case study

Family Hub outreach (Test, Learn and Grow programme)

In December 2024, the Cabinet Office launched a three-year programme (Test, Learn and Grow) to design, test and scale innovative solutions to complex challenges. The programme aims to close the gap between policy, delivery and users, and speed up cycles of learning and improvement, with multidisciplinary teams working alongside place-based delivery teams.

Before formally launching the programme, four 12-week 'demonstrator projects' were set up to stress-test the model of working. Two of these projects focused on improving family hub outreach – in partnership with Manchester City Council and Sheffield City Council – to increase the number of families accessing early years support.

Over 12 weeks, the multidisciplinary team worked closely with LAs and Family Hub delivery managers to understand the problem and barriers, identify relevant services, generate solutions, test prototypes, and implement data collection and feedback cycles. User research and engagement data provided insight on how the interventions were working in practice.

In Manchester, based on user research with Family Hub managers and staff, as well as a rapid evidence review, the focus was on testing pop-up outreach in community settings – meeting parents where they are, in libraries, GP surgeries, and food banks – to increase uptake of a language development screening called WellComm.

Feedback and data indicated that community spaces are an effective way to reach our target families. Early indicators suggest that 15 children signed up during the four-week testing period, with over half of these families being from the top 10% IMD decile [explain what this means]. Early signs also indicate that pop-up outreach works best when paired with an activity that hooks parents in, such as play sessions.

In Sheffield, parents attending a drop-in play workshop at the Family Hub were encouraged to enrol in a more intensive, evidence-based service called Making it REAL (Raising Early Attainment in Literacy). Based on community-specific barriers, two solutions were tested:

1. Demonstrating activities from the Making it REAL programme at play workshops.
2. Sending WhatsApp voice note testimonials from parents who had completed the programme in preferred languages.

The findings suggest that demonstrations may be an effective approach for increasing uptake of Making it REAL. Early indicators suggest that 27 parents signed up during the testing period, compared to just two parents in an unmatched comparison area. Audio messages showed weaker signals, but there is scope for future testing.

The timescales of this project limited the number and rigor of feedback loops. As a result, findings should be taken as an indication of early promise rather than firm solutions. Nonetheless the way of working showed real value – unifying teams across central and local government around a common problem, identifying and resolving data gaps to inform service delivery, and using information from users to make 'real time' changes to improve practice.

6.6. GROW: Robustly evaluating to inform decisions on potential expansion

There is a clear expectation across government that publicly funded programmes are robustly evaluated to assess whether they deliver their intended outcomes, provide value for money, and generate learning to inform future policy and spending decisions (see Magenta Book).

Iterative Test and Learn cycles help optimise individual components of an intervention, but they do not remove the need for robust evaluation of the intervention as a whole. Once a design has been refined through multiple 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 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 helps prepare interventions for this kind of robust evaluation in two key ways. First, by refining design in the Test phase, it stabilises delivery, clarifies mechanisms of change, and generates early evidence on feasibility and operational realities. Second, by involving analysts and researchers from the start, Test and Learn supports the development of roll-out plans that preserve the conditions needed for robust impact assessment – for example, through phased or randomised roll-out to enable counterfactual designs. Too often, government programmes are implemented at full scale before a stable model of delivery has been established, making later evaluation challenging or even impossible. Universal rollout removes opportunities to construct valid comparison groups and undermines the credibility of impact estimates. While evaluation planning should begin at the earliest stage of policy development, formal impact evaluation should only take place once an intervention is operating consistently and as intended. Test and Learn provides the bridge between early design and full-scale evaluation – ensuring that when a robust impact assessment is commissioned in the Grow phase, it is applied to an intervention that is both well understood and evaluable by design.

Co-design principles should also apply to evaluation planning. Bringing together policy teams, frontline delivery staff, analysts, researchers and technical specialists to co-design evaluation questions and methods ensures that evaluation is grounded in operational reality and produces findings that are understood and usable across the system.

Please note: Some interventions in complex and dynamic environments may need to use adaptive evaluation methods, such as Theory-based evaluation methods or developmental evaluation, throughout the life cycle of the policy. There is an annex to the Magenta Book on Handling Complexity in Policy Evaluation.

Checklist: Are you ready to move to Grow phase?

- ✓ **Key uncertainties and risks are resolved:** Risks and assumptions have been addressed, with evidence supporting the final approach.
- ✓ **Design is stable:** The intervention's key features, processes, and delivery model are clearly defined and not expected to change during the evaluation period.
- ✓ **Delivery is consistent:** Frontline staff, partners, and systems are delivering the intervention in a uniform way across locations or settings.
- ✓ **Sufficient scale and reach:** The intervention is operating with enough participants, locations, or activity to generate meaningful evaluation data.
- ✓ **Clear outcome measures:** Agreed outcomes, indicators, and data collection processes are in place and functioning reliably.
- ✓ **Resources and capacity confirmed:** Funding, staffing, and operational support are secured for the duration of the evaluation.
- ✓ **No major planned changes:** No significant policy, operational, or delivery changes are scheduled that would disrupt the intervention during evaluation.

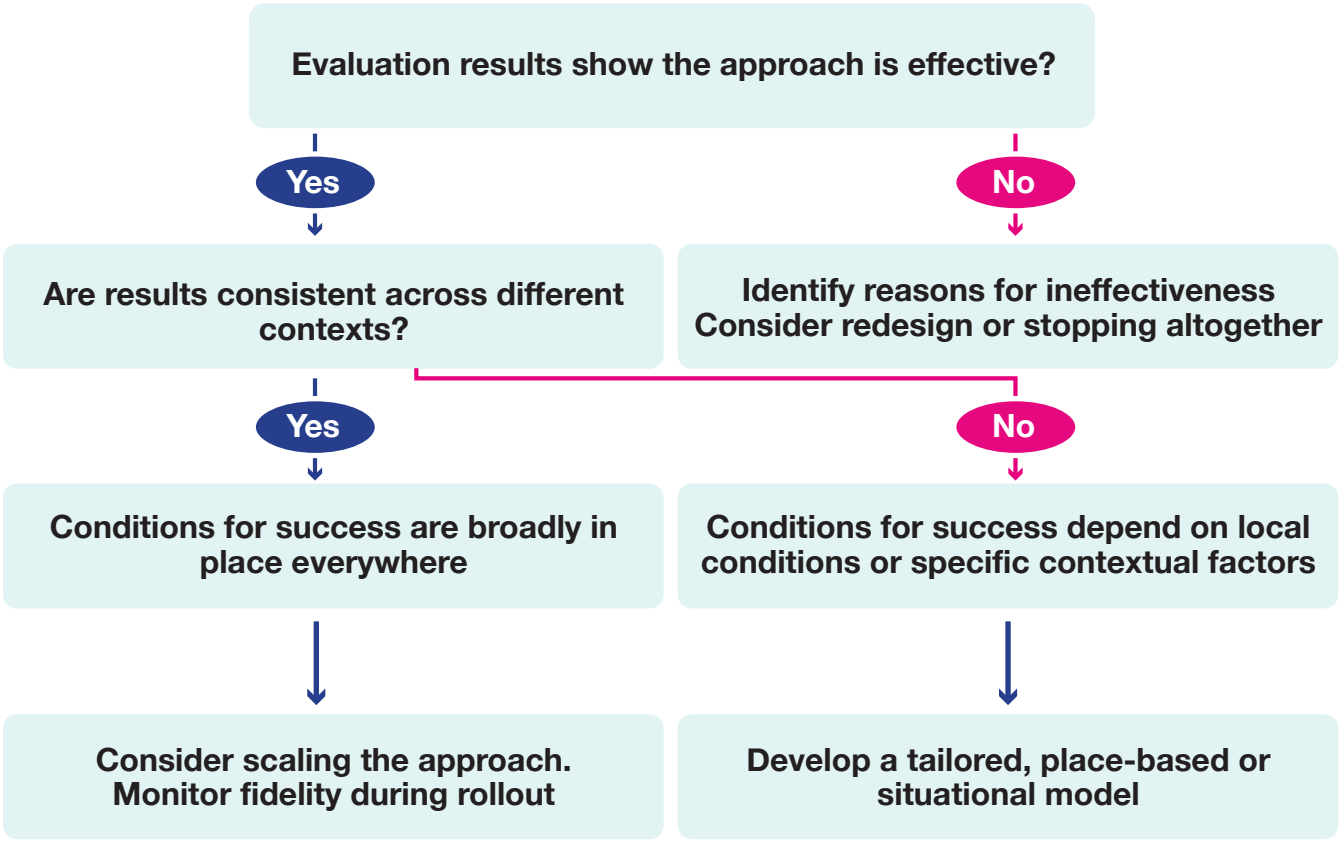
In the Grow phase, evaluation provides robust evidence on whether an approach achieves its intended outcomes under the conditions in which it is delivered. At this stage, evaluation is designed to assess impact using appropriate counterfactual approaches, such as randomised controlled trials or quasi-experimental designs.

Understanding the counterfactual means considering what would have happened if the intervention had not taken place. This allows evaluators to judge whether any changes can reasonably be attributed to the intervention, rather than other events or trends happening at the same time. By systematically collecting and analysing this evidence, evaluation can show not only if the approach works, but also for whom, under what circumstances, and at what cost.

Understanding if, for whom, how or why an approach has delivered impact is critical when considering whether an intervention should be scaled to a wider population or adapted for new settings. A process evaluation can reveal whether success depends on specific contextual factors – such as local infrastructure, workforce capacity, or community relationships – that may not exist elsewhere. If results show that outcomes are consistent across different contexts, this provides confidence that the intervention can be implemented more widely. However, if benefits are closely tied to local conditions, the evidence may point to the need for a more tailored, place-based model. In both cases, evaluation ensures that decisions about where and how to expand or adapt an intervention are based on credible evidence rather than assumptions, reducing the risk of

ineffective or inefficient rollout. As illustrated in Figure 6.7, the decision-making flowchart helps policymakers interpret evaluation findings and decide whether an intervention should be: (a) scaled more widely, (b) adapted to fit different local contexts, or (c) reconsidered altogether.

Figure 6.8: Decision pathway: Scale or localise?



Checklist: Practical steps for the Grow stage

Further methods guidance is available in the Magenta Book

- ✓ Establish a credible counterfactual (where feasible) – identify what would have happened in the absence of the intervention to assess whether observed changes can reasonably be attributed to it.
- ✓ Use process evaluation to understand context, identifying factors such as workforce capability, infrastructure, delivery partnerships or local culture that may influence effectiveness, and assessing whether success relies on conditions that may not exist elsewhere.
- ✓ Estimate value for money by considering both costs and benefits, proportionate to the scale and risk of the intervention.
- ✓ Decide the appropriate next step based on the evidence:
  - ✓ Scale the intervention where it demonstrates effectiveness across contexts
  - ✓ Adapt the model where benefits depend on local conditions
  - ✓ Reconsider the approach where evidence does not support further expansion

HM Treasury

Chapter Seven

Annex

7

Table 7.1 Cheat sheet

	What	How	Output
EXPLORE	Understand the system	Systems mapping/actor mapping/ behaviour change wheel.	A high-level map showing key elements, relationships and mechanisms in the system. Barriers and enablers influencing behaviour.
	Understand people's needs	User research, soft systems methodology.	Needs of people benefiting from a service or policy, needs of those delivering the service and the needs of wider stakeholders.
	Identify opportunities	Stakeholder workshops to identify bottlenecks, duplication/ inefficiency or behaviours that central to the problem and can be changed.	Identified areas in the system where a well-designed intervention could have a disproportionate impact.
	Decision point: What opportunity should we pursue first?		
CO-DESIGN	Generate ideas	Workshop with cross-functional team and partners to generate diverse ideas.	Long list of ideas.
	Design initial prototypes	Workshop to combine ideas into viable prototypes. Prioritise based on feasibility/impact. Get feedback.	Paper prototypes such as storyboards, mock-ups, or maps. Prioritisation criteria such as decision matrix or weighted scoring criteria.
	Test prototypes	Design quick experiments in small scale. Get qualitative feedback.	Actionable insight against agreed success criteria.
	Decision point: Is this the right opportunity to pursue? Which prototype should be taken forward? Which prototypes should be stopped?		
	Develop Theory of Change (ToC)	Workshop with diverse stakeholders.	Logic model with prioritised assumptions based on uncertainty and risk. Narrative of how change is intended to happen.
TEST	Identify most uncertain or riskiest parts of your ToC	Evidence base, insight from Explore and Co-design testing, professional judgment.	Activities, assumptions or mechanisms that need to be tested to inform next decision.
	Small scale evaluation on specific components	Any method from your research toolbox such as A/B testing, qualitative sprints, nimble trials.	Actionable insight against agreed success criteria.
	Decision point: How can we improve the intervention/policy? Do we need to re-test? Is this still the right opportunity to pursue? **Is the design, delivery and core processes stable and operating as intended?**		
GROW	Programme evaluation	Process, Impact and Value for Money Evaluations.	Publication.
	Decision point: Is the approach effective?		
	Scale or place based model	Monitor fidelity to key factors identified in evaluation.	Performance monitoring framework and data.

7.1. Useful resources

This section provides links to further information and useful resources about methods discussed within this annex.

Behaviour insights: 'Achieving behaviour change: a guide for national government' (published via Public Health England) uses the Behavioural Change Wheel Approach (Michie, Susan & Van Stralen, Maartje & West, Robert. (2011). The Behaviour Change Wheel: a new method for characterising and designing behaviour change interventions. Implementation science: IS. 6. 42. 10.1186/1748-5908-6-42).

Nesta: Nesta's *Test, Learn, Grow* playbook provides practical guidance on how to design and run small-scale experiments in public services, with a focus on testing assumptions, generating rapid feedback and building evidence for what works before scaling. Test and learn: a playbook for mission-driven government | Nesta

Public design case study bank: This collects real-world examples from across government showcasing how design practices – such as user research, prototyping, systems mapping and Co-Design – have been applied to policy and service challenges, offering practical insights and lessons for how design can support better outcomes. Public Design Evidence Review: Case Study Bank (HTML) - GOV.UK

Soft systems methodology (SSM): This is a systems thinking method designed for tackling 'messy' social problems in which different stakeholders hold different views about what the problem is and how to resolve it. Rather than assuming a single, clearly defined problem, SSM helps facilitate structured debate, explore multiple worldviews and model alternative interventions using tools like rich pictures, root definitions and conceptual models. Soft Systems Methodology

Service Design: Service design methods offer valuable tools for the Explore stage, as they help teams build a shared, evidence-based understanding of complex systems and the people within them. Method Library – This is Service Design Doing

Systems thinking: The Government Office for Science's toolkit, *Systems Thinking for Civil Servants* encourages teams to "map out a rough outline of the stakeholders within a system, or list possible enablers and blockers to success" even in just ten minutes. Systems mapping enables teams to visualise how different actors, behaviours, infrastructures, and institutional dynamics interconnect to create the current policy landscape. Systems thinking for civil servants – GOV.UK

Theory of Change: If you have a role in the planning, implementation or evaluation of an intervention, whether that is a project, programme, policy or government service, you should consider adopting a Theory of Change approach. The Theory of Change Process – Guidance for Outcome Delivery Plans – Government Analysis Function