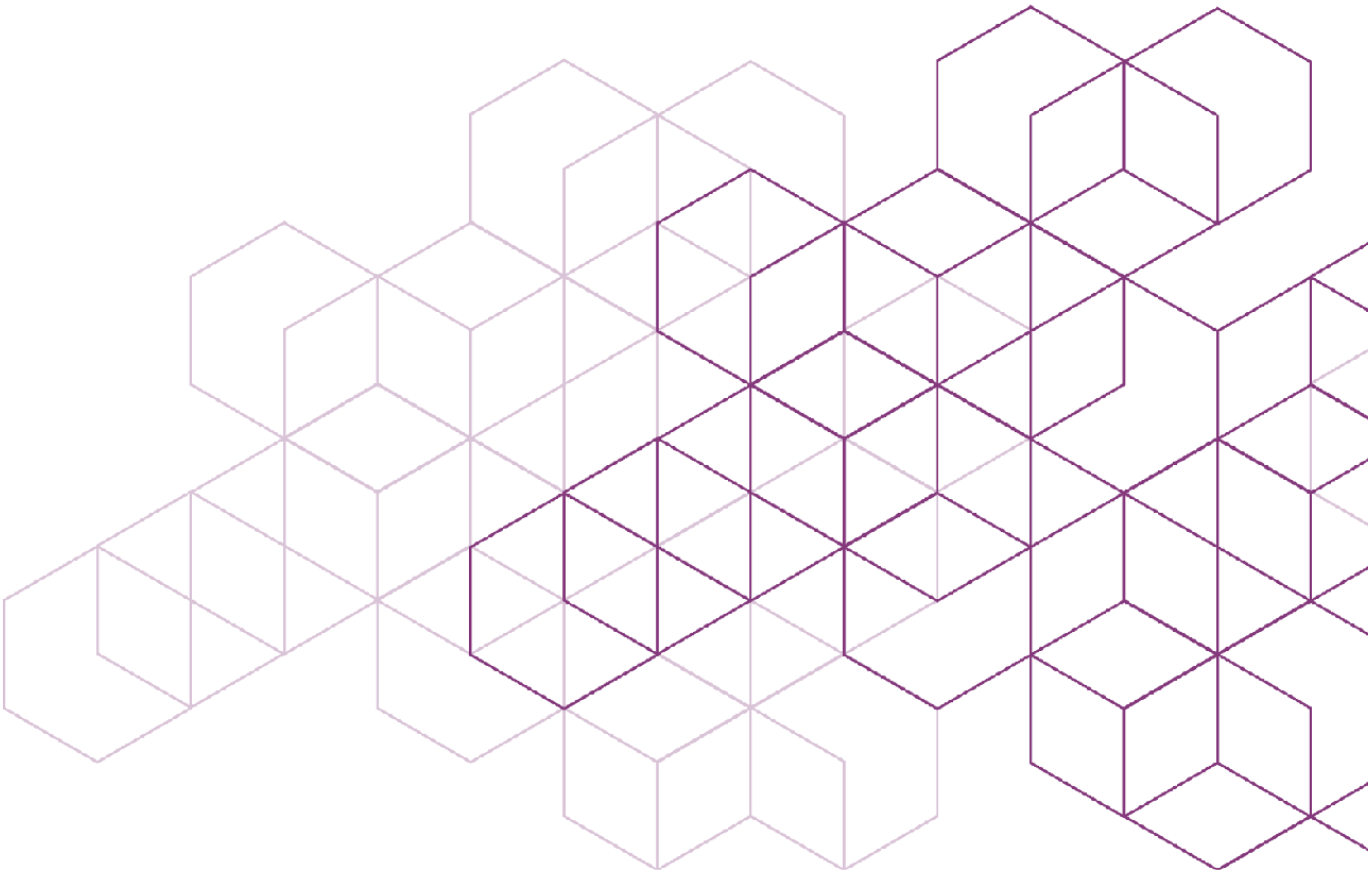
Government
Skills

Evaluating Management and Leadership Training in the Civil Service

A how-to guide to enable rigorous and feasible evaluations

Christian Schuster
Ine Steenmans



Government
Skills

Executive summary

This report provides a how-to guide for evaluating management and leadership training programmes, drawing on lessons learned from evaluations of four management and leadership Civil Service training programmes. The guide supports both evaluation specialists and training programme designers, providing practical guidance on embedding evaluation in programme design and implementation. It also aims to support the dialogue between designers and evaluators, recognising that high-quality evaluation requires close collaboration between these groups from design through delivery to analysis.

The how-to guide sets out a credible, rigorous and feasible approach to provide insights into whether training programmes are effective, for whom they are effective and how programmes could be further improved. The suggested approach is not the only possible evaluation design, and this level of evaluation may not be proportionate for all training programmes – some programmes may be small-scale or not sufficiently strategically important to warrant this intensive approach. However, the guide does provide a tried-and-tested method to evaluate management and leadership training programmes in the Civil Service in a rigorous and informative way. The approach moves beyond reaction surveys, which measure participant satisfaction but not actual learning or effectiveness. Evaluators and programme teams should adapt the methods suggested in this guide proportionately to their specific context and resources.

Key principles underpinning this approach:

- **Begin with comprehensive programme mapping:** Understanding what training actually delivers - not just what it claims - provides the foundation for measuring the right outcomes.
- **Pre-post surveys are essential for measuring effectiveness:** Post-only satisfaction surveys cannot capture whether participants learned or changed their practice.
- **Combine surveys with qualitative methods:** Interviews and observations explain how and why training achieves effects, informing actionable recommendations.
- **Apply proportionality:** Match evaluation intensity to programme scale, maturity, and decision needs.

The evaluation approach comprises the following steps (summarised in Figure 1).

Programme mapping and theory of change development

In a first step, the curriculum and training content is mapped in detail (gathering all programme materials to document what content is actually taught, how sessions are structured, and what outcomes are expected). Based on a detailed understanding of training content, a theory of change for the training can be developed. This theory of change needs to be specific enough to serve as a foundation for measurement – for instance, by identifying the specific management practices the training seeks to improve among participants.

Pre-post participant surveys

Based on this theory of change, a pre- and post-programme survey with training participants can be designed. The survey should cover – among others – knowledge (through objective knowledge questions), attitudes and practices in the management and leadership domains the training covers. The how-to guide explains this approach, and also both provides advice on effective question writing and proposes an approach to survey implementation which enables high response rates in both the pre-programme and post-programme survey. Pre-post survey results provide insights into training effectiveness, for instance by helping understand whether participants improve their knowledge of management domains covered in the training and change their management behaviour in response to training. Post-survey results also shed light on training engagement with different training components, which can provide one potential explanation for the effectiveness patterns observed.

Interviews, focus groups and observations

To complement survey results and develop recommendations to improve effectiveness, the guide provides practical approaches for planning and conducting interviews, focus groups, and observations of training delivery. The combination of pre–post survey data and qualitative insights can provide programme leads with comprehensive understanding of how and why training achieves (or does not achieve) changes.

Optional extensions

Finally, the guide discusses extensions evaluators may wish to consider to further enhance the robustness and scope of the evaluation.

Randomised control trials can provide programme leads with more robust causal insights into training effects, and the how-to guide details a tried-and-tested approach for implementing them for management training programmes in the Civil Service. The successful use of an RCT in the Achieving Your Potential evaluation – alongside other recent Government examples including the Digital Excellence Programme and the Evaluation Academy – demonstrates that RCTs are feasible within normal Civil Service operational contexts. Programme leads are therefore encouraged to consider whether an RCT design is appropriate before defaulting to a purely observational approach.

The guide also introduces two approaches to understand downstream training effects: surveys with team members and, where accessible, data from the Civil Service People Survey can shed light on training effects on the direct reports of trained managers; and administrative data analyses can help explain the broader workforce and productivity impacts of training.

Figure 1. Evaluation pathway showing core approach and optional extensions. Each component is detailed in subsequent sections of this guide.

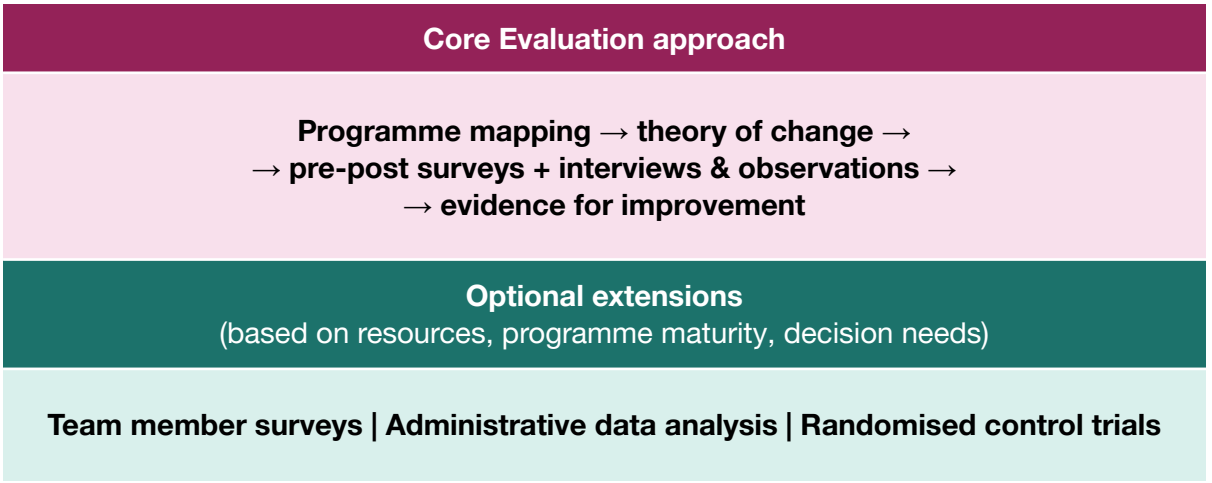


Table of contents

Introduction

The quality of management and leadership in an organisation is a key predictor of organisational productivity (Bloom et al., 2013). An effective Civil Service thus requires effective management and leadership, at all levels of hierarchy. How can organisations enhance the quality of management and leadership? Management and leadership skills training programmes are one important answer. A recent meta-analysis finds that "despite much heterogeneity in their design, management skills training programs are usually effective at improving outcomes." (Busso et al. 2023). That management programmes are effective, on average, however, does not imply that every management and leadership training is effective (Lacerenza et al. 2017).Even where training programmes are well-designed, for instance, "limited cognitive capacity and task overload in the aftermath of leadership training programs can reduce the transfer of learning into action" (Haunstrup and Jensen, 2024, p. 684).

The relationship between training and practice change is therefore more complex than a simple linear progression from training input to improved performance. Factors as diverse as cognitive load from workplace responsibilities, line manager support, and organisational culture mediate whether and how learning translates into changed behaviour. For evaluation purposes, measuring whether practice change occurs and exploring the barriers and enablers to transfer provides actionable insights for programme improvement, even without fully explaining all causal mechanisms at play.

Determining whether a given training is effective – and investments into training represent value for money in the Civil Service and for taxpayers – thus requires evaluating training. Evaluations can shed light on key questions, such as:

- What are the characteristics of civil servants participating in the training, and does this match the programme's target audience?
- How satisfied and engaged are participants with the training programme and its components?
- Do training participants improve their knowledge of the content covered in the training?
- Is the training effective at changing the attitudes, networks and/or management practices of participants?
- Does management training participation affect the career trajectories of participants or their direct reports?
- Does management training participation increase the productivity of organisational units led by training participants?
- Why does the training attain learning or practice change for some of the topics covered, but not others?
- How can ineffective training components be improved?

Evaluations, of course, presuppose prior training design. The approaches in this guide help answer questions about training effectiveness (whether it works), mechanisms (how and why it works), and improvement (how to strengthen it). They cannot rigorously resolve what

content civil servants should be trained in in the first place, nor questions about optimal training formats or mode (online vs. in-person) without comparative evaluation designs. For such design questions, formative research and needs assessments provide more appropriate starting points.

Traditionally, in the Civil Service and elsewhere, training is evaluated through a questionnaire administered at the end of the training (Kirkpatrick and Kirkpatrick, 2009). Such questionnaires typically measure how satisfied participants were with the training and instructor and other training-related attitudes and perceptions, such as self-assessments of competencies taught. End-of-training reaction surveys provide helpful early warning systems: if participants find training irrelevant, poorly organised, or unengaging, they are unlikely to apply it later (Kirkpatrick and Kirkpatrick, 2009). Reaction surveys can also collect helpful information to improve training design.

However, surveys do not provide insights into whether training is effective. Satisfaction with training often does not correlate with actual learning (Hughes et al., 2016; Uttl et al., 2017). Similarly, self-assessments of competencies often do not correlate with actual competencies, including in the public sector (Mejia-Guerra et al., 2023). Training satisfaction is also typically only weakly correlated with application of material – i.e. behavioural change of participants as a result of the training (Gessler, 2009). Evaluating the effectiveness of training thus requires more than reaction surveys.

This how-to guide provides a tried-and-tested method in the Civil Service to evaluate the effectiveness of management and leadership training. It supports two audiences: evaluation specialists conducting assessments, and programme designers and delivery teams who need to embed evaluation in their programmes. The guide aims to strengthen dialogue between these groups by clarifying what makes training evaluable and how evaluation can inform programme improvement.

This guide was developed and trialled in evaluations of four different management training programmes led by the Cabinet Office. The approach proceeds in five steps. First, the training curriculum is mapped. Second, a theory of change is developed which reflects the training programme's content and is specific enough to inform qualitative and quantitative measurement. Third, pre- and post-surveys with training participants are designed and implemented in a way that secures high response rates and provides evidence for learning (competency and skill improvement of participants), network building of participants, changes in attitudes and practice change from the training (the extent to which training participants manage or lead differently). Fourth, interviews and participant observations are designed and implemented to understand and complement survey results. Qualitative insights can shed light on how and why training achieves (or not) changes, and how training design could be improved to further effectiveness, for instance.

These four steps enable a step-change in the quality of evidence available in Civil Service departments about the effectiveness of management training and, analogously, other training programmes, in particular on leadership. In a final step of the how-to guide, we introduce extensions to these steps, subject to feasibility: randomised control trials to

enable more robust causal inferences; surveys with team members and data from the Civil Service People Survey or departmental employee surveys (such as pulse surveys) to understand training effects on participant direct reports and to measure management practice change more robustly; and administrative data analyses to understand broader workforce and productivity impacts of training. Table 1 summarises the different data sources and the questions they shed light on.

Table 1. Summary of evaluation data sources and the insights they provide

Data source	What it can measure (examples)	What it cannot measure or provide (examples)
End-of-programme training reaction surveys	• Training satisfaction • Engagement with training components • Characteristics of civil servants selecting into programme • Short open text comments from participants on how to improve training	• Training effectiveness, and its causes • Learning of participants • Analyses of who benefits most from the training in terms of learning and behavioural change
Pre- & post-programme surveys	• Learning of participants • Training effectiveness in terms of attitudinal, networks and management practice change of participants • Plus, all measures typically included in an end-of-programme reaction survey (see above) • Causal evidence on the effects of training on participant outcomes (when combined with a suitable control or comparison design, such as a randomised control trial)	• Training effects on workforce outcomes, teams or productivity • Unexpected explanations for (lack of) training effectiveness (factors not included in survey) • Ideas on how to improve ineffective training components beyond survey respondent comments
Interviews	• How and why training achieves (or fails to) change through participants' lived experiences and sense-making • Specific barriers and enablers to applying learning in workplace contexts • Unexpected factors influencing effectiveness	• Generalised patterns across large participant populations • Quantitative measurement of the magnitude of learning or behaviour change • Objective verification of self-reported practice changes

	• Actionable recommendation for improvement based on participant insights	
Training observations	• Significance of training delivery features, including facilitator actions, participant engagement patterns, group dynamics • Whether programme design is implemented as intended • Factors in learning environment that support or hinder learning and engagement	• Individual internal experiences, motivations, comprehension beyond observable behaviour • Whether participants retain learning or apply skills after training concludes • Effectiveness of learning activities between training sessions
Dedicated team members surveys, or data from the Civil Service People Survey or existing departmental surveys	• Effects of training managers on their direct reports: are they more motivated, willing to stay and do they rate the quality of management of their line manager more highly?	• Training effects on workforce outcomes of team members (such as retention)
Workforce records	• Training effects on participant careers (for instance, does their performance rating and promotion trajectory improve?) • Training effects on team member workforce outcomes (for instance, does turnover and sick leave decrease?)	• Training effects on organisational productivity (rather than individual performance)

A worked example in this guide: the ‘Fundamentals’ line manager training

To illustrate key concepts in practice, this guide uses as an illustrative example a fictional Fundamentals line manager training programme designed for newly appointed Civil Service managers. This example is fictionalised to enable broader application across different departments and contexts, while drawing on structural features common to Civil Service line management programmes (such as Foundation and Practitioner). This 4-month programme combines in-person workshops, self-led learning, and peer action learning sets to develop delegation and feedback skills.

Developing a theory of change that is measurable

Key actions

- ✔ Gather all training materials (curriculum, exercises, communications) to understand what content is actually delivered
- ✔ Create a programme map documenting goals, participants, activities, outputs, outcomes, and assumptions
- ✔ Distinguish between intended outcomes (aspirational claims) and expected outcomes (what content realistically teaches)
- ✔ Develop a theory of change articulating how and why training is expected to produce change
- ✔ Identify which data sources can measure each outcome level in your theory of change

Map before you measure

Before collecting any data, evaluators need to understand what the training is, what it is trying to achieve, and how it is meant to work. Programme mapping is a key tool for this purpose. Programme mapping gathers the materials used in training delivery, clarifies the learner journey, and surfaces the intended outcomes and assumptions. Programme mapping protects against two common pitfalls: evaluating only easily measured outcomes such as satisfaction ratings rather than learning or behaviour change and mistaking aspirational claims of learning outcomes for the actual content and impact of learning.

In practice, mapping is proportionate. The Magenta Book encourages evaluators to scope first and match effort to decision needs and constraints (UK Government, 2020). A light-touch inventory may therefore be sufficient for a short training programme, while a multi-module, mixed-mode offer will likely require a more extensive mapping to provide the foundation on which a testable theory of change can be built. The programme mapping approach described here is compatible with other programme description frameworks such as the TIDieR (Template for Intervention Description and Replication) checklist, but provides training-specific guidance on capturing pedagogical features and learning outcomes central to training evaluation. Ideally, this mapping and theory of change development already occurs at the programme design stage. This fosters a clear, shared understanding of how the programme should work and ensures programme design reflects a clearly delineated and detailed theory of change.

Materials to gather during evaluation scoping

The process often begins with collating all the programme materials that reveal what was actually delivered and what resources participants had access to. Collating these programme materials serves dual purposes: it provides data for assessing quality and alignment of content with programme goals and participant needs, as well as helps formulate interview questions and interpret results. As a minimum, these should include:

- **Curriculum and training materials:** Check whether learning objectives are stated, and review all materials (slides, workbooks, exercises) to identify what content is actually delivered (which may differ from stated objectives).
- **Exercises and assignments:** Note the learning activities and any follow-up task requirements.
- **Participant outputs:** If relevant, request samples of project briefs or reflective outputs that show how participants enact learning.
- **Communications:** Materials such as participant joining instructions, emails and online portals give insights into how the training is framed to learners, what outcomes are promised, and what kinds of commitment is expected (for example, pre-work and time between sessions).
- **Feedback forms:** End-of-session and end-of-programme feedback can reveal issues that may be relevant to investigate further.

Build a clear and reusable programme map

The output of mapping should be a clear, structured description of what the training programme is: its goals, participants, activities, outputs, intended outcomes and underlying assumptions. This programme map then informs development of the theory of change, which articulates **how** and **why** the programme is actually expected to produce change (and not just how programme documentation states the programme intends to produce change, as in the map). The map can take various forms (a table, structured document, or visual diagram). What matters is that the map can be read and used by colleagues, project partners and stakeholders throughout the evaluation. Table 2 lists elements that should be captured in any map.

Table 2. Key programme elements

Element	Description	Illustrative Fundamentals example
Ethos and goals	Summarise the training objectives (both stated and actual, see section below) and core programme problem statement.	Equip newly appointed line managers with foundational people management capabilities through practical application. Key objectives: developing delegation skills, constructive feedback conversations, and inclusive team environments.

Participants	Profile who the training is for, and on what basis they are selected.	Civil servants newly appointed to line management (within 12 months), selected via departmental nomination or self-referral. 20-25 per cohort.
Activities and inputs	Identify all formal sessions, e-learning, coaching, peer groups, and any informal elements learners are encouraged to use.	Four modules over 4 months: 1) delegation; 2) feedback and coaching; 3) performance management; 4) building inclusive teams. Between modules, participants complete workbook exercises and join action learning sets (peer groups of 5-6) to discuss implementation. Informal WhatsApp groups provide ongoing support.
Immediate outputs	Capture all tangible outputs participants create (action plans, quiz responses, reflections) and practice opportunities created.	Completion of all four modules, submission of one-year action plans identifying three practices to implement, workbook exercises, attendance at action learning sets.
Training outcomes	Identify intended learning outcomes (knowledge, skills, attitudes), behaviour change outcomes (management and leadership practices on the job), organisational impacts (for example, team climate, service performance).	Learning: knowledge of delegation and feedback frameworks, confidence to hold difficult conversations, peer support networks. Behaviour change: increased frequency of structured one-to-ones, appropriate delegation patterns. Organisational impact: team members report clearer expectations and greater manager support.
Assumptions, context, uncertainties	Identify what else must be true for participants to apply learning in their work, for example, time to practise, permission to experiment, line-manager support, and access to resources. Record uncertainties to clarify or investigate later.	Significant assumptions are that participants have line manager permission to experiment in their current roles and their workload allows time for additional coaching practice. A key question to explore is whether Modules 3-4 (2 hours each) provide sufficient time for in-depth coverage of performance management principles, or whether content needs rebalancing across modules.

Indicative evidence	For each outcome, note potential data sources and measurement approaches. Keep this high level, as detailed instrument design comes later.	Potentially measure with knowledge quizzes, confidence ratings, and network formation in pre–post surveys, while 3-month 360-degree feedback from team members could capture behaviour change.

Engage stakeholders during mapping

Interviews with delivery and commissioning teams during the mapping phase surface tacit design logic and implementation realities that formal documentation rarely captures. These conversations help understand what makes programmes distinctive and how they work in practice, shaping realistic evaluation scope.

Evaluation effectiveness benefits significantly from early collaboration with programme teams. Where external delivery partners are involved, specify evaluation requirements (data access, consent procedures, observation permissions etc) in contracts at the procurement stage. Agree data collection responsibilities and establish communication channels during programme mapping to ensure evaluation activities integrate smoothly with delivery. This collaborative foundation positions evaluation as supporting programme improvement rather than external audit.

Testing programme assumptions

The programme map identifies two types of assumptions that require testing during evaluation. Programme logic assumptions are intrinsic to the training's theory of change (for example, “participants will change their management behaviour after learning new frameworks”). These are tested through standard evaluation data collection such as surveys measuring behaviour change and interviews exploring whether and how participants applied learning.

Contextual assumptions concern the conditions necessary for training to work (for example, “participants have line manager support to experiment” or “workload allows time for practice”). These underlying assumptions are equally vital to test but require explicit attention in data collection design. Contextual assumptions can be tested through multiple data collection methods, each offering different insights. Interview questions should probe these conditions directly (for example, “What support did you receive from your line manager to try new approaches?” or “What barriers prevented you from applying what you learned?”). Observations can reveal whether promised practice time actually materialises. Surveys can also measure contextual assumptions through structured questions – for instance, rating scales for line manager support or quantifying available practice time between sessions. Where contextual assumptions prove unfounded, even well-designed training will struggle to achieve intended effects. Identifying this through evaluation enables programme redesign or better participant targeting.

Distinguish intended, expected, and observed training outcomes

Examining programme documentation as part of the programme map (Table 2) helps identify the intended outcomes of a programme at different levels: learning, behaviour change, and organisational impact. However, these intended programme outcomes often differ from what content realistically can be expected to deliver, and from the outcomes that are actually observed in any evaluation. Distinguishing intended, expected, and observed outcomes is essential for a valid evaluation.

Table 3. Comparing intended, expected and observed outcomes for fundamentals

Outcome type	**Definition**	**Illustrative example with Fundamentals**
Intended	What programme documentation claims training will deliver	"Develop confident, capable line managers who effectively lead teams and drive performance" → ambitious, but vague
Expected	What curriculum analysis reveals the training content actually covers, based on material depth, practice opportunities, and learner readiness	• Knowledge of feedback frameworks • Skills in structuring feedback conversations • Increased confidence to hold difficult conversations • Expanded peer support networks to reflect on challenging team conversation scenarios
Observed	What evaluation evidence eventually shows occurred; this is what the evaluation primarily examines and measures	To be determined via evaluation data collection (for example, surveys, interviews)

Large gaps between intended and expected outcomes indicate the need for programme redesign before evaluation proceeds. For example, if programme documentation claims to "develop expert performance management skills" but curriculum analysis reveals only two hours of introductory content, the programme requires content strengthening or outcome recalibration before evaluation investment is effective. Where intended and expected outcomes align reasonably well, evaluation focuses primarily on measuring expected outcomes (what training content realistically covers) against observed outcomes (what

actually changed). This approach ensures evaluation measures what training actually teaches rather than aspirational claims.

Extracting expected outcomes from programme content requires systematically coding learning activities using established taxonomies such as Bloom's (knowledge, comprehension, application, analysis, synthesis, evaluation) rather than accepting stated, intended objectives. Systematic, bottom-up coding of learning activities reveals what materials actually teach. This distinction between outcome types serves two purposes in evaluation. First, it enables diagnostic assessment: a programme might meet expected outcomes given training content yet fall short of ambitious intended claims, suggesting opportunities to strengthen programme design or recalibrate outcome claims rather than indicating poor delivery. Second, it focuses instrument design by ensuring evaluation measures what the training actually covers rather than aspirational objectives that exceed realistic content scope.

From programme map to theory of change

A theory of change (sometimes abbreviated as ToC) provides a credible, context-sensitive explanation about how the training leads to outcomes. It articulates the programme's underlying logic: what activities occur, what changes these produce, and why those changes happen. For readers unfamiliar with theories of change, the Magenta Book provides guidance on ToC development across government programmes (UK Government, 2020). In training evaluation specifically, a theory of change links activities to change through explicit assumptions about necessary conditions and causal mechanisms, forming the evaluation's conceptual foundation.

Ground your theory of change in tested frameworks

Evaluators may find it helpful to structure their thinking about training outcomes using established frameworks. Two commonly used, potentially complementary, frameworks are introduced below – Kirkpatrick's Four Levels and COM-B – but these are not requirements, many other frameworks exist, and the Kirkpatrick and COM-B frameworks are used in this guide as examples of how frameworks can be used rather than as recommended frameworks. What matters is articulating clear, testable assumptions about how training produces change. Evaluators should select or adapt frameworks that best fit their programme logic or develop theories of change without relying on any particular framework.

The first, **Kirkpatrick's Four Levels,** provide a linear progression from participant experience to organisational results:

- **Level 1 - Reaction:** Did participants find the training relevant and engaging?
- **Level 2 - Learning:** What knowledge, skills or attitudes changed?
- **Level 3 - Behaviour:** What changed in day-to-day practice?
- **Level 4 - Results:** What contribution to team or organisational outcomes is plausible?

Table 4. Example outcomes and indicators for Kirkpatrick's Four Levels for Fundamentals

Kirkpatrick Level	**Example outcomes identification for Fundamentals**	**Example potential measurement indicators for Fundamentals**
Level 1 - Reaction: Did participants find the training relevant and engaging?	Satisfaction with training; perceived relevance to role	• End-of-programme satisfaction ratings • Open-ended feedback on most/least useful modules
Level 2 - Learning: What knowledge, skills or attitudes changed?	Knowledge: feedback models, delegation framework; Skills: structuring difficult conversations; Attitudes: confidence for difficult conversations	• Pre-post knowledge quiz on delegation frameworks • Observed application of feedback structure in role-play • Pre-post confidence self-rating
Level 3 - Behaviour: What changed in day-to-day practice?	Quality of feedback given; appropriate delegation of task	• Self-reported frequency of structured one-to-ones • 360-degree feedback from team members on feedback quality • Self-reported delegation frequency
Level 4 - Results: What contribution to team or organisational outcomes is plausible?	Team member experiences of line manager support	• People Survey items for participants' teams on manager support and work satisfaction

Kirkpatrick's levels provide one way to structure outcome categories and remind evaluators to look beyond learner satisfaction (Level 1). However, this framework assumes a linear progression from training to results that does not reflect current understanding of how learning occurs. In particular, the move from learning (Level 2) to behaviour (Level 3) is more complex than the sequential approach in Kirkpatrick suggests, and is mediated by factors including workplace context, cognitive load, and organisational culture. For these reasons, while Kirkpatrick's levels can be a useful scaffold for identifying outcome types to measure, evaluators should not treat them as an endorsed model of how training works. The Kirkpatrick framework's value lies in structuring measurement, not explaining mechanisms. It also reminds us that low levels of observed behaviour change can reflect context constraints and not only content quality.

A second framework, **COM-B** offers an alternative lens, unpacks behaviour change into **C**apability, **O**pportunity, **M**otivation → **B**ehaviour. For line-management training, COM-B prompts diagnostic questions: What **capabilities** does the course build that learners can

demonstrate during or immediately after? What **opportunities** must be present back at work for transfer to occur? What **motivations** are strengthened, such as confidence or commitment to a coaching style?

Kirkpatrick and COM-B can serve complementary purposes in evaluation design. Kirkpatrick's levels can help structure what outcomes to measure across the causal chain (reaction, learning, behaviour, results), providing a measurement scaffold. COM-B can provide a diagnostic lens for interpreting findings: when surveys show capability increased but behaviour did not, COM-B directs investigation toward opportunity barriers or motivation gaps. Used together, these frameworks can strengthen both instrument design and interpretation, though neither is required for rigorous evaluation.

Articulate causal mechanisms through if-then-because statements

Frameworks such as Kirkpatrick and COM-B provide useful scaffolding for organising outcome types and identifying components of behaviour change. However, these frameworks do not themselves explain the causal mechanisms that are the specific processes through which training produces change. To develop this clarity, evaluators need to articulate explicit causal chains showing how and why training is expected to work in their specific context.

If-then-because statements provide a format for making these causal chains explicit and testable. This format forces explicit causal reasoning about how and why training produces change, making mechanisms (the generative processes through which training creates effects) visible and testable.

The basic form is an if-then-because statement:

> ***If*** learners undertake [specific activities] in [this format],
> ***then*** specific [learning] and [behaviour changes in management practice] and [downstream outcomes] will occur,
> ***because*** [mechanisms such as practice with feedback, social learning, or identity reframing] operate in [named context conditions].

For complex, multi-module programmes, you may need several if-then-because statements, one per major causal pathway or mechanism clusters, rather than attempting to capture all programme logic in a single narrative.

Fundamentals examples of two if-then-because statements for a theory of change

Causal pathway 1: learning

> *If* participants engage in four modules with structured practice, peer feedback, and action learning sets,
>
> *then* capability (framework knowledge, coaching skills) and confidence increase,
>
> *because* deliberate practice with feedback builds self-efficacy.

Example measurement implications: this learning pathway suggests pre–post surveys could capture whether capability (e.g. framework knowledge gains and confidence) increases, while interviews could explore how deliberate practice with peer feedback builds self-efficacy.

Causal pathway 2: application

If participants have line manager permission and support, and protected time to apply learning,

then behaviour change occurs (structured feedback, delegation),

because peer support networks and action plans bridge intention to implementation.

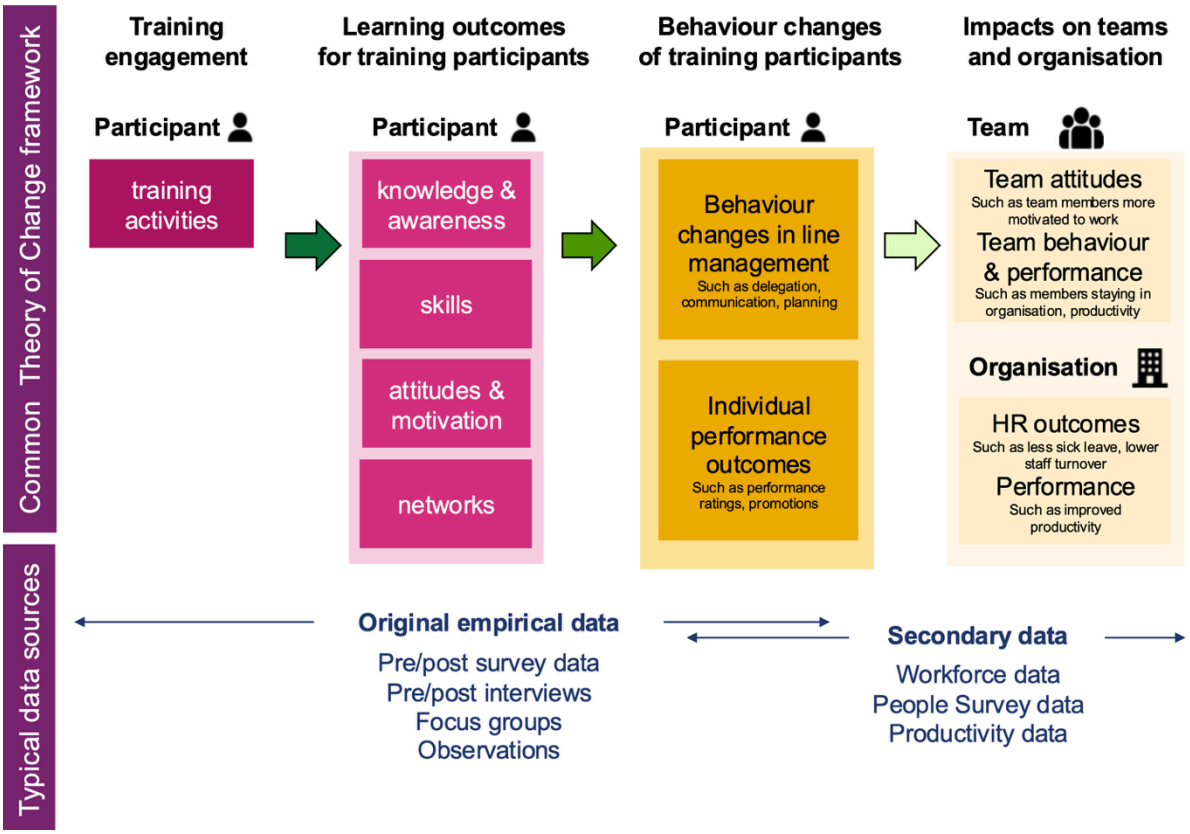
Example measurement implications: This application pathway indicates need for interview questions probing contextual factors (line manager support, protected time) and suggests value of 360-degree feedback from team members to measure behaviour change from multiple perspectives.

Facilitation and commissioning teams often hold tacit knowledge about how components hang together. Discussions with programme designers, facilitators and managing teams, potentially supported with a focused workshop to review the draft ToC typically improves accuracy and legitimacy. An evaluation's theory of change is a living object that develops iteratively, which the team records and revisits as new data and analysis becomes available.

From theory of change to measurement strategy

A complete theory of change maps not only outcomes across the causal chain but also identifies which data sources can measure each level. Figure 2 illustrates the typical data sources for training evaluation.

Figure 2. Evaluation data sources across the training-to-impact pathway.



A full evaluation across all relevant areas (training engagement, learning outcomes, behaviour changes, and impacts on team and organisation) requires multiple complementary data sources:

- **Original empirical data** (pre–post surveys, interviews, focus groups, and observations) capture individual-level learning outcomes (knowledge, skills, attitudes, networks) and participant-reported behaviour changes. These data sources and collection methods provide insight on what effects were generated and how and why programme mechanisms operate.
- **Workforce records** enable measurement of individual performance outcomes, including participants' career progression (promotions, performance ratings, retention) and their direct reports' workforce outcomes (retention, sick leave, performance).
- **Secondary data sources** (People Survey team-level results, and productivity metrics) enable measurement of team and organisational impacts where available, though attribution typically becomes more challenging at these downstream levels.

The following sections detail practical approaches for designing and implementing data collection methods to gather evidence from each data source.

Pre- and post-surveys with participants

Key actions
✔ Design a survey questionnaire which maps on and measures key components of the programme's theory of change ✔ Keep questionnaire length to under 20 minutes ✔ Pre-test survey questions through cognitive interviews ✔ Incentivise high response rates, for instance by conditioning programme access on pre-programme survey completion and generating personalised diagnostics for post-programme survey respondents ✔ Ensure sufficient statistical power for analysis by focusing pre- and post-survey implementation on training programmes with 500 or more participants (presuming a pre-post response rate of 80% is achieved)

Pre- and post-surveys are surveys with participants before the training programme and at the end of – or after – the training programme. An individual identifier in the survey (for example, an email address inserted at the end of the survey, or a unique survey link emailed to participants) allows analysts to link individual responses to the pre-survey with the same individual's post-survey response.[1] As a result, participants' attitudes and management practices before and after the training can be compared, providing insights into whether those attitudes and practices shifted with training participation, and for whom such shifts occurred.

What surveys can(not) measure in a theory of change

As mentioned in the introduction, most existing training evaluations in the Civil Service conduct post (reaction) surveys with participants, to gauge training satisfaction. The proposed pre- and post survey approach retains this element: post surveys can and should contain reaction survey questions. However, pre- and post surveys also enable analysts to measure other parts of the theory of change, beyond training engagement, in particular:

- **Learning:** Do participants have greater knowledge about the material taught in the training?
- **Attitudinal change:** Do participants change their attitudes towards management or leadership – for instance, the extent to which they care or enjoy being managers?
- **Network change:** Do learners change their reported peer networks to support their development?

[1] Care needs to be taken to ensure there is explicit informed consent by participants for this linkage.

- **Behavioural change:** Do participants manage or lead differently?

Behavioural change – participants changing their management practices – is the most downstream impact of training in the theory of change which pre- and post-surveys with training participants can plausibly measure. Understanding further downstream impacts – for instance on the motivation of team members of the manager taking the training, or the performance of a unit or organisation – requires other data sources (see the 'Extensions' section of this guide). To illustrate the limits of training participant surveys for these downstream impacts: self-reported performance in employee surveys is weakly correlated with actual performance, in part as individuals are too lenient when self-assessing their performance (Heidemeier and Moser, 2009).

Designing effective survey questionnaires

Survey questionnaires need to be short enough – typically 20 minutes or less – to safeguard adequate response rates and quality (see Revilla and Höhne, 2020). At the same time, questionnaires need to be sufficiently detailed and specific to help evaluators understand training effects in different parts of the theory of change.

There is ample guidance on how to write effective questionnaires which ensure that survey questions accurately measure constructs, such as managerial practices or attitudes (Wolf et al., 2016). A first step in any questionnaire development thereby lies in defining the concepts for measurement along the theory of change:

- Which concepts or content might participants have learned in the training? An example question would be a multiple-choice question on what 'systems thinking' is.
- Which particular attitudes towards management might the training programme have shifted? An example would be a Likert agreement scale question on the extent to which participants enjoy being line managers.
- Which networks of participants might the training programme have enhanced? An example would be a Likert agreement scale question on the extent to which participants have peer networks to source ideas to improve their management.
- Which particular management practices might participants have changed as a result of training participation? An example would be a frequency scale question about the frequency with which participants discuss with individual team members what motivates or demotivates them.

Defining these questions requires close attention to training content module-by-module, to ensure the evaluation does not measure outcomes which training is not designed to achieve. Where possible, questionnaire designers should draw on existing, validated survey questions. Validated survey questions are items that have been tested and refined through prior research to ensure they reliably measure the intended concept, produce consistent results across different populations, and minimize bias or misinterpretation. Drawing on validated survey scales lowers the pre-testing burden (see below) and strengthens the

credibility and comparability of findings. Using established scales also helps align results with broader evidence bases and reduces risks of bias or measurement error from untested items.

While otherwise preferable, existing scales may not be appropriate for every evaluation context. Training programmes might cover content for which validated scales do not exist, or for which validated scales do not meaningfully cover the behavioural or attitudinal changes a training programme seeks to attain. Where new questions need to be developed, they need to be designed for (see Wolf et al., 2016):

- Comprehension, including by asking one question at a time, using familiar, specific and simple words, avoiding double negatives and using as few words as possible
- Retrieval, accounting for how respondents recall information and not asking respondents to retrieve information they are unlikely to remember (for example, due to long recall periods)
- Reporting, including through easy-to-understand response scales (for example using scales with 5 to 7 scale points and clear labels), and questions which minimize social desirability bias (a tendency for respondents to respond in ways that make themselves look good). Particularly questions in which participants assess themselves – for instance, the frequency with which they engage in specific management practices – risk suffering from social desirability bias. Careful wording to minimize this risk is thus warranted, for instance by using specific and time-bound questions, such as "'In the past four weeks, how many one-to-one meetings did you hold with each direct report?"
- Variation, which means differences in responses may be expected between training participants and within-participants over time, and responses are not heavily skewed (for example, almost all participants responding "strongly agree"). Lack of variation complicates assessing training effects.

Next to substantive questions to evaluate training effects, questionnaires should contain questions about the demographic and professional background of participants. Such questions can help descriptively understand who enrols in the programme. For instance, civil servants with disabilities or from ethnic minority backgrounds might be underrepresented. Demographic and professional background questions can also help understand heterogeneous training programme effects. By way of example, they can help understand whether training effects are larger for civil servants in higher ranks or for civil servants who have previously not taken any other management training.

Prior to fielding, survey questionnaires should be pre-tested through cognitive interviews. This step involves evaluators probing how respondents understand, interpret, and mentally process each question, ensuring that wording is clear, concepts are consistently understood, and response options align with intended meanings. Ideal cognitive interview partners include previous training participants (or, where the training is not yet online, civil servants who would be eligible for the training).

An example questionnaire to illustrate and facilitate replication of the pre-post survey approach in other training evaluations can be found in Appendix A.

Securing high response rates and sufficient response numbers

End-of-training reaction surveys often suffer from poor response rates. Even when dedicated time is allocated to surveys in the last training session, response rates in management training programmes may remain well below 50%. Low response rates both undermine statistical power for analysis and, more importantly, raise concerns about representativeness. Less satisfied participants may also be less inclined to spend time on a training survey, for instance (see Cardell, 2022). In a pre–post survey design, low response rates are particularly problematic. Only participants who complete **both** the baseline and endline surveys can be included in the pre–post analysis.[2] Consequently, attrition at each stage compounds. By way of example, if 50% of participants respond at baseline and 50% at endline, and these response patterns are independent, the effective sample is 25%. This puts a premium on a careful and incentivised approach to securing higher response rates in both the pre- and post-programme surveys.

One tried-and-tested approach in the Civil Service – applied in evaluations of the Foundation, Practitioner and Senior Practitioner management training programmes – incentivises high response rates as follows:

- Pre-programme survey: participants are required to complete a pre-programme survey to access online learning materials (though participants can omit responding to individual questions to ensure substantive responses to questions remain voluntary). This is operationalised by requiring a four-digit code to access online learning materials. Participants completing the pre-programme survey and inserting their email address at the end of the survey are automatically emailed this code. Participant email addresses are then also used to link pre- and post-programme responses of participants.
- End-of-programme survey: participants are required to complete an end-of-programme survey to receive a personalised management diagnostic. Participants are provided with time to complete the survey in the final training session. Participants are then asked to discuss their personalised management diagnostic results with their trainer and peers in the session. To be able to participate in that discussion, participants thus need to complete the survey, incentivising high response rates. The management diagnostic is customised based on survey responses. It benchmarks how frequently a given participant applies management practices taught in the training programme compared to peers (other training participants, drawing on data collected in the pre-programme survey). Participants insert their email address at the end of the survey and automatically

[2] The post-programme survey should nonetheless be conducted among all participants. Comparisons of responses of those completing only the post-programme versus those completing both (pre and post) can help assess the representativeness of those who respond to both (pre and post) surveys.

receive this diagnostic.[3] Email addresses are also used to link participant responses to the pre-programme survey.

Next to high response rates to ensure representative results, a sufficient total number of responses is needed for statistical power to detect effects. Statistical power is the ability of an evaluation to detect a real training effect when one truly exists or, in other words, that an evaluation will correctly identify a training effect rather than miss it and conclude "no effect" when there is one. As a rule of thumb for pre–post survey analysis, just under 400 matched respondents are needed in a pre-post survey approach to be able to detect relatively small training effects (d = 0.2) with 80% power (assuming a pre–post correlation of r = 0.5). It is important to be able to, statistically speaking, detect 'relatively small' effect sizes. Training programmes can enhance the quality and frequency of good management practice, but programmes often take place in a context in which participants typically already engage to some extent in these practices – thus meaning training generally tends to have effect sizes that are described as "small" (but this does not mean they are necessarily not practically meaningful and important).

The proposed approach thus requires training programmes of sufficient size to avoid concerns about lack of statistical power. Even with relatively high response rates – for example, a pre–post response rate of 80% - upwards of 500 training participants are needed for meaningful analysis. Sample size requirements further increase if evaluations are also supposed to assess the extent to which training is differentially effective for different groups of participants – such as participants from different departments, demographic backgrounds or participation modes (for example, online compared with. in-person).

Deep dive for technical specialists only: analysing data and reporting results

Programme evaluation through pre- and post-programme surveys requires survey data analysis skills. Programme designers may thus need to bring in data analysts to support analysing survey data and reporting results. How to statistically analyse survey data is well documented elsewhere, providing insights into issues such as weighting, item nonresponse, and other sources of error (Wolf et al., 2016a). Survey-based training evaluations should draw on this body of work. Beyond ensuring a valid analysis, evaluators may wish to focus on:

- Pre-specifying (latent) measurement in a pre-analysis plan. During questionnaire development, state which items capture observable constructs (for example, number of workshops attended) and which capture latent constructs (such as, quality of performance management). For latent constructs, pre-specify the

[3] In the evaluation of the Foundation, Practitioner and Senior Practitioner training programmes, the diagnostic was coded in HTML and SmartSurvey's automatic email trigger functionality was used to email out diagnostics after surveys are completed.

latent-variable method (such as CFA/IRT), scoring rules, model-fit criteria, and any decision thresholds.

- Making the analysis fully reproducible and auto-updating. Script all data cleaning and analyses so results can be rerun end-to-end. If the sample grows as training continues, automate reporting (such as with R Markdown) so tables and figures update automatically when new observations are added.

Interviews and observations of training

Key actions

- ✔ Design interviews, focus groups, and observations as complementary methods that triangulate evidence
- ✔ Plan interview timing strategically: some immediately post-training (fresh impressions), some at 2-6 months (sustained application)
- ✔ Derive interview questions from the programme's theory of change and survey findings to build coherent evidence
- ✔ Consider interviewing approaches that explicitly test programme theory assumptions with participants
- ✔ For observations, provide advance notice, explain purpose, and maintain transparent but unobtrusive presence
- ✔ Begin analysis during fieldwork, coding emerging themes and grading evidence by confidence level

Interviews, focus groups, and observations are among the most common instruments for gathering lived experience data in training evaluation. They provide individual perspectives on learning processes, explore group-level sensemaking of shared and diverging experiences, and reveal delivery dynamics and early indicators of learning (for example, participants express a change in ideas) that cohort-level data, such as that collected in surveys, cannot easily capture. In line with the Magenta Book encouragement for proportionate data collection, a compact data collection plan with a handful of interviews may suffice for interventions with straightforward intervention mechanisms, while complex multi-component programmes warrant richer designs using multiple methods to engage with diverse stakeholder perspectives and experiences.

Specialist expertise considerations: While programme teams can conduct basic interviews with appropriate preparation, some methods, such as some observations, require specialist evaluation skills. Evaluators should consider whether to engage Government Social Research (GSR) support or external evaluation specialists for more complex qualitative data collection.

Design complementary methods from inception

Training evaluation interviews, focus groups, and observations each contribute distinct but complementary evidence: interviews surface individual mechanisms and contextual factors; focus groups enable dialogue about converging and diverging experiences; observations

reveal implementation variations and learning dynamics. Their complementarity further strengthens validity where corroboration across sources is achieved.

Data collection activities can be staggered to align with different analysis stages. Early observations and initial interviews help verify programme delivery and surface assumptions. Mid-programme interviews support emerging theory development about how change occurs. Later interviews and focus groups refine findings and test provisional conclusions. Different data collection activities can inform one another iteratively where these are sequenced carefully to build on each other.

Plan interviews to explore lived experience

Interviews provide depth on individual learning journeys, application challenges, and the factors that enabled or hindered transfer. They are typically semi-structured, balancing specificity with openness and flexibility to follow participant narratives with consistency to enable cross-case comparison. Interviews typically run 30-60 minutes, and should start by ensuring confidentiality and anonymisation are addressed.

Who to interview, and when

Interview sampling typically aims to capture heterogeneity of learning experiences rather than representativeness of participant profiles with attributes such as grade or organisation. Where feasible, interviews with learners several months or even years post-training are undertaken to assess sustained behaviour change. Facilitator interviews reveal implementation variations and delivery challenges. Line managers or team members of participants can validate whether workplace context enabled application, though access may be limited.

Interview timing is chosen strategically. Immediate post-training interviews capture fresh impressions and intended actions, while interviews at 2 to 6 months assess genuine application once participants have had opportunity to implement learning in their work contexts. Conducting interviews after post-programme survey results are available enables exploration of unexpected patterns of strong/weak effects identified in survey data.

What to ask

Interview prompts and questions are derived from the programme's theory of change, and other data such as survey findings. This ensures interview data builds towards a coherent evidence base for validity and inference strength. Questions are generally open-ended to avoid leading participants or introducing bias with too narrow or biased questions.

An example interview guide to illustrate interviewing approach in training evaluations is included as Appendix B. Effective generic training evaluation interview questions include:

- **On significant experiences:** What has been a significant experience or change while participating in the training?

- **On enablers and constraints:** What has enabled or challenged your application of learning?
- **On grounding with examples:** Can you give an example of when you used something from the programme?
- **On exploring propositional insights:** What could have made the training more effective?

Testing programme theory with participants

Evaluators may also adopt interviewing approaches that explicitly test the programme's theory of change with participants (Manzano, 2016). This technique is drawn from realist evaluation, an evaluation approach that focuses on understanding what works, for whom, in what circumstances, and why (see Magenta Book, UK Government 2020, for comprehensive guidance). Rather than only asking open-ended questions to understand participant experience (as detailed above), this approach presents participants more directly with the programme's assumed causal mechanisms and asks them to validate, challenge, or refine these assumptions. For example: “The programme assumes that action learning sets help bridge the gap between knowing what to do and actually doing it. Has that been your experience? Why or why not?” This approach can reveal whether the mechanisms evaluators assumed are operating actually match participants’ lived experiences, strengthening the validity of causal claims.

Use focus groups to explore shared sensemaking

Focus groups expand individual perspectives into collective dialogue, exploring how peers interpret training experiences together. They reveal shared norms and divergent views that individual interviews may not surface. Group sizes typically comprise 4-7 individuals, and it can be helpful to sample participants from the same cohorts to provide a shared reference frame. Sessions of 45-60 minutes work well, structured around key themes such as expectations versus reality, persistent application, and improvement recommendations. Evaluators should consult established focus group guidance for detailed facilitation techniques, particularly regarding question design to avoid leading questions, managing group dynamics to ensure all voices are heard, and supporting participants to develop shared understanding whilst respecting divergent perspectives.

Observe training delivery transparently

Observation is less commonly used in Civil Service evaluation than interviews or surveys but provides unique insights into training delivery dynamics that other methods cannot capture. Observation requires specific expertise: developing robust observation protocols, maintaining objectivity during sessions, and systematically recording complex interactions demand training and practice. Evaluators new to observation should seek support from

badged Government Social Researchers or specialist guidance to develop appropriate instruments and approaches.

Observation reveals delivery dynamics, participant engagement, and implementation variations. The focus of the observation guides will vary by key objectives for the observation. Where an evaluation examines how learning influences behaviour, guides often focus on understanding how different facilitation approaches and participant interactions support and reinforce learning processes. Training observers often pay particular interest to content alignment, engagement patterns, practice opportunities, and facilitator adaptations.

In sensitive line management and leadership training contexts, it is important to provide advance notice of any observation and explain your purpose (evaluating the programme, not individuals). Observers always seek to adopt transparent but unobtrusive presence and can seek layered consent for session observation as well as breakout session participation.

Observation structure and data recording

Observation approaches range from highly systematic to more exploratory. Systematic structured observation uses predetermined observation frameworks with specific behaviours or interactions to record (for example, counting how many times facilitators use open questions, timing different activity types, recording which participants speak). This approach enables quantitative analysis and comparison across sessions but may miss unexpected dynamics. Less structured observation begins with broad focus areas and allows observers to note whatever seems significant as it unfolds. The choice depends on evaluation context: systematic observation works when you know what matters and want to measure it consistently; exploratory observation suits programmes where key delivery factors are still being identified.

Data recording for observations also varies. Structured observation sheets with pre-defined categories (such as rating scales for engagement levels, checklists of facilitation techniques observed) provide consistency but constrain what gets captured. Field notes with descriptive narratives capture richer contextual detail but require more analysis time and observer skill. Video or audio recording enables detailed post-session analysis but may affect participant behaviour and raises consent requirements. A practical middle ground combines a simple observation framework (key areas to focus on) with open field notes, allowing both structured comparison and flexibility to capture unexpected insights. Where multiple training sessions are observed, developing an observation framework after the first session and applying it systematically to subsequent sessions can balance exploration with rigour. Appendix C provides an example observation protocol used in Civil Service management training evaluations that demonstrates this middle-ground approach, organising observations thematically with attention to theoretically salient delivery features while remaining open to emergent patterns.

Analyse data iteratively throughout fieldwork

Analysis can be commenced during data collection, coding emerging themes and patterns as they surface (namely, systematically identifying and labelling recurring themes in interview transcripts or observation notes).

To manage explorations of inferences about programme contributions while acknowledging uncertainty at different stages in the evaluation, it can be helpful to grade evidence by confidence level. One approach uses three categories: **validated** (multiple-source confirmation), **emerging** (some supporting evidence), **preliminary** (single-source). These categories are illustrative; evaluators should adapt confidence grading to their context and data sources available. Evidence grading approaches can also be useful to analyse interview, focus group, and observation data independently before comparing with survey results. This preserves inductive integrity (allowing themes to emerge from data rather than imposing predetermined categories), then enables systematic **triangulation** (comparison across different data sources). Triangulation identifies **convergence** (where sources agree), **contradiction** (where sources disagree), and **complementarity** (where sources provide different but compatible insights). Programme designers working with evaluation specialists can contribute valuable context during this triangulation process, helping interpret findings in light of programme realities.

Extensions: team member surveys, administrative data analysis and randomised control trials

Key optional actions to extend the analysis further:

- ✔ Conduct surveys of team members of managers participating in training programmes or draw on data from existing employee surveys with team-level results – such as the Civil Service People Survey – to understand training effects on the motivation and wellbeing of team members of training participants
- ✔ Draw on administrative data – workforce records and case productivity records – to evaluate training effects on workforce and productivity outcomes
- ✔ Conduct randomised control trials of training programmes – for instance, by randomly assigning interested participants to earlier and later programme starts – to provide more robust, causal estimates of training effects

Pre-post surveys and qualitative data collection can provide rich insights into training engagement, the effects of training on participant learning, attitudes and practices, and mechanisms through which the training is (in)effective. These approaches are less effective at capturing outcomes further downstream in the theory of change—such as participants' workforce performance, the attitudes and behaviours of their team members (for example, work motivation), and the productivity of their units or organisations. We therefore consider potential extensions to the pre–post participant survey and qualitative data collection methods to better measure these downstream effects and enable more robust causal identification of training impacts.

Understanding training effects on team members through employee surveys and the Civil Service People Survey

Management quality is a strong predictor of employee motivation, well-being, and retention (Meyer-Sahling et al., 2018). A core downstream expectation of management training is therefore teams with greater motivation and well-being. Understanding this downstream impact requires surveying subordinates of training participants. This often presents logistical and incentive challenges: how to identify team members, reach them and convince them to participate in a pre- and post-survey relating to their manager. Surveys of employees of participating managers thus often suffer from lower response rates than surveys of training participants themselves (Haunstrup and Jensen, 2024). They also add a costly (in terms of effort and survey design) evaluation layer. Where practically feasible, however, they do enhance rigor of the training evaluation: they measure downstream impacts on teams and enhance the robustness of inferences about training effects on

management practice. Employee ratings of management practices of their superior provide an important complement to pre–post self-assessments of managers (Fleenor et al., 2010).

Where surveys of team members (direct reports) of training participants are infeasible, team-level results from the Civil Service People Survey are a valuable alternative data source. The Civil Service People Survey generates over 12,000 team-level reports every year, including on key employee attitudes in the team (for example, motivation or wellbeing) and perceptions of the quality of management of the direct superior (Cabinet Office, 2022). With due confidentiality considerations – including safeguards to ensure results for particular teams are not shared publicly – team-level People Survey results can thus be drawn on to understand whether employee attitudes and perceptions of management in teams led by training participants improve after the training. This also offers the advantage of drawing on measures from a widely known and recognised existing measurement instrument. For this analysis to be viable, several important conditions need to be met:

- Several hundred training participants each lead a team for whom Civil Service People Survey team-level results are available
- Team-level identifiers in the Civil Service People Survey can be matched to team or unit identifiers of training participants (this linkage is complicated in departments in which the list of units in the People Survey does not match the list of units in workforce records)
- Internal analysts have or can gain access to Civil Service People Survey team-level results over time
- Existing Civil Service People Survey questions provide relevant measures for the effects on participant team members which may be expected given the content of a particular training

Understanding training effects on workforce and productivity outcomes through administrative data

Access to two types of administrative data allows evaluators to understand further downstream impacts of training: workforce records and productivity records.

Workforce records help evaluators understand, in the first place, whether training shapes workforce outcomes for participants. For instance, are they more likely get promoted after training, receive higher performance ratings, remain in the organisation and report less sick leave? Workforce records of training participants and of a comparable (control) group of non-training participants enable estimation of the difference-in-difference in outcomes of the two groups before and after training. This difference-in-difference approach compares how much the outcomes change over time for the training group relative to how much they change for the non-training group, isolating the effect of the training by taking into account background trends that affect both groups.

Where workforce records identify subordinates of training participants, the analysis can be extended to workforce outcomes of team members. Are subordinates of managers receiving training – relative to a control group of subordinates in the same organisation and

with similar characteristics but of managers not receiving training – more likely to remain in the team, receive higher performance ratings, obtain promotions and report fewer sick days?

As detailed in Schuster (2026), extending the analysis to workforce records is not without challenges. To cite just three examples: where training records are not integrated with workforce records, and hence there is not an exact way to match people in both datasets, (fuzzy) matching is required, to link those registering for training in one dataset with those same individuals in workforce records (such as via email addresses or names, given the absence of a Civil Service ID). This process requires close attention to data protection. The analysis also requires consistent records over time, to trace training impacts over time and construct a control group with similar pre-training trends in outcomes such as promotions or performance ratings. Yet, changes to Human Resource Management Information Systems (HRMIS) can complicate constructing consistent panels of individuals over time. Finally, treatment effects of individual training programmes on outcomes such as promotions or retention are likely small even in the case of otherwise successful programmes, as a range of factors unrelated to training shape those outcomes (such as budget availability for promotions). Large samples traced over several years are thus required. Extending the analysis to workforce records thus requires prior consideration of the availability of records and their accessibility, also from a data protection perspective.

Administrative productivity records can extend the analysis further downstream, to assess training effects on organisational or unit productivity. The digitisation of government operations has multiplied the number of potential data sources available to analysts to understand productivity in government (Rogger and Schuster, 2023). In particular, many government organisations now process administrative cases digitally. By way of example, the Department of Work and Pensions processes benefit claims; the Home Office asylum claims; and HMRC tax claims, among others. These digital case records can be repurposed into objective measures of performance of teams – for instance, by measuring the speed with which similar cases are processed and the quality of case processing (Best, Fenizia and Khan, 2023). Similarly, frontline service delivery organisations have digitised operations and metrics, which can be repurposed into productivity measures – for instance, in health, education, the police or prison service (Andrews et al., 2023).

As a result of the potentially available data, records can be drawn on to estimate productivity impacts of training. Do teams whose managers receive training become more productive compared to a control group of teams with similar characteristics in the same department but whose managers do not receive training? As greater productivity is often a key ulterior objective for investing in management training, extending the analysis to productivity records provides decision-makers with a key metric to gauge value for money of a training programme. The scope conditions for conducting this analysis are significantly narrower than in for the pre–post survey approach, however:

- The training includes several hundred participants who manage teams in either the same department or the same function (such as procurement)

- The same administrative case or service delivery records are available for teams of trained managers and a control group (such as participants all manage teams processing asylum claims, or all manage teams delivering similar health services or all manage teams procuring goods and services)
- Administrative case records are consistent and available over time, at least a year before and after the training.

Providing causal estimates on training effects through randomised control trials

The pre–post survey approach and its extensions with administrative data all rely on observational inferences. Training effects are estimated by comparing participants' responses before and after the training (pre–post survey) or by comparing the workforce outcomes or productivity of participants and their teams before and after the training relative to a matched control group (administrative data analysis).

What, however, if managers who are most motivated to improve their management practices select into the training? A pre–post comparison might find improvements in management practices of participants. Those improvements, however, might not stem from training, but from the motivation of participants to improve their management more generally, which might lead them to engage in other concurrent activities to improve their management (for example, by observing their superiors, or reading about management). Vice versa, what if departments intentionally send managers whose management practices are on a downward trajectory into training? A pre–post comparison might find no improvements in management practices. Yet, in the absence of training, management practices might have deteriorated. These instances are examples of the so-called endogeneity problem: selection into the training is shaped by factors (such as the motivation of managers) which also influence the development of management practices over time of participants. In such instances, an observational approach to inference (surveying only a treatment group of those receiving the training) would provide biased estimates of training effects.

By contrast, randomised control trials (RCTs) enable causal and unbiased estimates of training effects. Randomised control trials randomly assign participants into treatment and control groups. Treatment group participants receive the training; control group participants do not. Survey data – and, where feasible, administrative data – is collected from both treatment and control group participants before and after the treatment group training. The difference-in-difference in outcomes between two groups before and after the training provides unbiased estimates of training effects.

Drawing on the recent RCT of the 'Achieving your Potential'[4] management training programme in the Civil Service, several lessons to enhance the practical feasibility of RCTs of management and leadership training programmes in the Civil Service can be stipulated:

- A wait-listed experimental design can enhance the acceptability of an RCT to participants as all participants eventually participate in the programme, only at different times. In a wait-listed design, treatment group participants are randomly assigned to participate in the training earlier, and control group participants assigned to participate later, after the completion of the treatment group training and the endline measurement.
- Careful consideration should be given to designing effective incentives for completing both the pre- and post-surveys among treatment and control group participants. Baseline survey completion can be encouraged by making it a prerequisite for registering in the training course (prior to random assignment). Endline survey completion can similarly be incentivised by requiring it to begin the course (for the control group) or to receive the training certificate (for the treatment group). As described in the pre–post survey approach, the endline survey can also be programmed to generate a personalised management diagnostic, automatically emailed to participants upon completion. This diagnostic can serve as an additional incentive: for the treatment group, it may be discussed in the final session, and for the control group, in the first session of the training.

RCTs can enable rigorous testing of pilot programmes, prior to their expansion, to inform whether and how to roll out the pilot, for instance by shedding light on which pilot training content is effective at achieving learning and behavioural change and which not.[5] New programmes typically also offer greater flexibility in terms of participant selection and wait-listed randomisation can thus more easily be built into their design. By contrast, RCTs of established programmes can help assess whether mature training programmes which have been refined over years are, in fact, effective at achieving intended training outcomes. This typically requires a change in how participants are selected for the programmes, and thus close engagement with departments booking places on to the programmes.

Before committing to an RCT design, evaluation teams should consider several key factors to assess feasibility. First, a suitable control or comparison group must be available – for example, a waiting list of participants not yet enrolled in training, or a group receiving alternative support without the specific intervention under evaluation. Second, the programme must have sufficient enrolment to meet minimum statistical power requirements. Thus, the practical and ethical implications of withholding or delaying training for a control group should be considered, ensuring this is consistent with programme objectives and departmental responsibility. Fourth, reliable data collection through, for

[4] Other recent examples of RCTs in Cabinet Office skills programmes include the Digital Excellence Programme evaluation (Cabinet Office, 2025a) and the Evaluation Academy waitlisted RCT (Cabinet Office, 2025b), both providing further evidence of the feasibility of experimental design within Civil Service operational contexts.

[5] Though such pilot programmes need to have sufficient training participants to meet the statistical power constraints for analysis described above.

example, pre- and post-survey completion by both treatment and control participants, is essential to the validity of the design. Where these conditions are met, RCTs offer strong design for generating causal evidence on training impact.

Conclusions

This guide has presented a comprehensive, adaptable approach to evaluating management and leadership training in the Civil Service. Starting with programme mapping and theory of change development, proceeding through pre–post participant surveys and qualitative data collection, and extending where feasible to team member surveys, administrative data, and randomised control trials, the approach can be scaled proportionately to match intervention complexity and available resources.

Effective training evaluation in practice is underpinned by common principles. First, an effective evaluation is grounded in an explicit theory of change that provides the foundation for measurement design and informs programme redesigns and transfer of evaluation learning to other programmes. Second, the integration of multiple methods such as surveys, interviews, focus groups, and observations enables triangulation across data sources and strengthens evaluation validity. Third, collaborative consideration of evaluation assumptions with programme delivery teams and wider stakeholders enhances the accuracy and the legitimacy of findings. Finally, the prioritisation of measurement for understanding whether training achieves its intended effects on learning, behaviour change, and downstream outcomes over measurement of what is easily accessible (such as end-of-training satisfaction) is essential to impactful evaluations.

The evaluation approach presented in this how-to guide provides a set of options that can be adapted depending on a programme's stage of development and strategic importance. This level of evaluation is not proportionate for all training programmes – small-scale programmes or those not strategically important may be suited to lighter-touch approaches. For programmes where comprehensive evaluation is appropriate, methods can be applied iteratively: programmes can be evaluated – for instance, via pre–post surveys and interviews – until sufficient data for a credible and well-powered analysis is secured. Based on the analysis and recommendations, programme designers can work to improve the programme. Once major changes have been implemented, another evaluation may be worthwhile to shed light on the effectiveness of the changes made and inform future changes. For established programmes showing consistent effectiveness, evaluation can transition to lighter-touch monitoring rather than intensive data collection.

The evaluation approaches in this guide enable Civil Service departments to answer the fundamental questions: does this training work, for whom, and how can it be improved? Evaluators should begin with programme mapping and theory of change development, then select methods proportionate to their programme's scale and maturity. For many programmes, well-designed pre–post surveys combined with interviews provide sufficient evidence for improvement. Extensions such as administrative data analysis or randomised control trials can be added where decisions warrant additional investment in causal rigour. Robust evaluation ultimately serves programme improvement – strengthening both individual training programmes and, over time, management and leadership capability across government.

References

Andrews, K., Kim, G., Rogers, H., Sharma, J. and Venegas Marin, S. (2023) Government Analytics Using Measures of Service Delivery. In Rogger, D. and Schuster, C. (Eds.) *The Government Analytics Handbook: Leveraging Data to Strengthen Public Administration*, Washington DC: World Bank, p. 327-344.

Best, M., Fenizia, A. and Khan, A. (2023) Government Analytics Using Administrative Case Data. In Rogger, D. and Schuster, C. (Eds.) *The Government Analytics Handbook: Leveraging Data to Strengthen Public Administration*, Washington DC: World Bank, p. 327-344.

Bloom, N., Eifert, B., Mahajan, A., McKenzie, D., Roberts, J. (2013) Does Management Matter? Evidence from India. *The Quarterly Journal of Economics*, 128:1, p. 1-51.

Busso, M., Park, K. and Irazoque, N. (2023) The Effectiveness of Management Training Programs: A Meta-Analytic Review. *IDB Publications (Working Papers) 12782*, Inter-American Development Bank.

Cabinet Office (2022) Civil Service People Survey 2021: Technical Guide. Retrieved from https://www.gov.uk/government/publications/civil-service-people-survey-2021-results/civil-service-people-survey-2021-technical-guide

Cabinet Office (2025a) Evaluation of the Digital Excellence Programme (DEP) for senior leaders: a pilot randomised control trial (RCT). Cabinet Office, Government Skills. Retrieved from https://www.gov.uk/government/publications/evaluation-of-the-digital-excellence-programme-for-senior-leaders

Cabinet Office (2025b) Evaluation Academy Final Evaluation Report. Evaluation Task Force, Cabinet Office and HM Treasury. Retrieved from: https://www.gov.uk/government/publications/evaluation-academy-final-evaluation-report/evaluation-academy-final-evaluation-report-2025-html

Cardell, C., Yuce, T., Zhan, T., Eng, J., Cheung, E., Etkin, C., Amortegui, D., Jones, A., Buyske, J., Bilimoria, K., Hu, Y. (2022) What They Are Not Telling Us: Analysis of Nonresponders on a National Survey of Resident Well-Being. *Annals of Open Surgery*. 12, 3(4), p.1-7.

Fleenor, J., Smither, J., Atwater, L., Braddy, P., Sturm, R. (2010) Self–other rating agreement in leadership: A review. *The Leadership Quarterly*, 21(6), p. 1005-1034.

Gessler, M. (2009) The correlation of participant satisfaction, learning success and learning transfer: an empirical investigation of correlation assumptions in Kirkpatrick's four-level model. *International Journal of Management in Education*, 3(3-4), p. 346-358.

Haunstrup, J., & Jensen, U. (2024) Leadership training and just-in-time nudges: A field experiment on the transfer of learning into action. *International Public Management Journal*, *27*(5), p. 684-708.

Heidemeier, H. and Moser, K. (2009) Self-other agreement in job performance ratings: a meta-analytic test of a process model. *Journal of Applied Psychology*, 94(2), p. 353-370.

Hughes, A. M., Gregory, M. E., Joseph, D. L., Sonesh, S. C., Marlow, S. L., Lacerenza, C. N., Benishek, L. E., King, H. B., Salas, E. (2016) Saving lives: A meta-analysis of team training in healthcare. *Journal of Applied Psychology,* 101(9), p. 1266-1304.

Kirkpatrick, D. and Kirkpatrick, J. (2009) Evaluating: Part of a ten-step process. In *Evaluating Training Programs*. Berrett-Koehler Publishers: San Francisco, USA, p. 3-20.

Lacerenza, C., Reyes, D., Marlow, S., Joseph, D., & Salas, E. (2017) Leadership training design, delivery, and implementation: A meta-analysis. *Journal of Applied Psychology, 102*(12), p. 1686-1718.

Manzano, A. (2016) The Craft of Interviewing in Realist Evaluation. *Evaluation*. *22*(3), p. 342-360.

Mayne, J. (2012) Contribution analysis: Coming of age? *Evaluation*. 18(3), p. 270-291.

Mejía-Guerra, J., Schuster, C., Rojas Wettig, M., Sass Mikkelsen, K., & Meyer-Sahling, J. (2023) *Making National Statistical Offices Work Better: Evidence from a Survey of 13,300 National Statistical Office (NSO) Employees in 14 Latin American and Caribbean Countries*. Inter-American Development Bank.

Meyer-Sahling, J.-H., Schuster, C., and Sass Mikkelsen, K. (2018) Civil service management in developing countries: what works? Evidence from a survey with 23,000 civil servants in Africa, Asia, Eastern Europe and Latin America. Report for the UK Department for International Development (DFID).

Revilla, M., and Höhne, J. K. (2020) How long do respondents think online surveys should be? New evidence from two online panels in Germany. *International Journal of Market Research*, *62*(5), p. 538-545.

Rogger, D. and Schuster, C. (2023) *The Government Analytics Handbook: Leveraging Data to Strengthen Public Administration*, Washington DC: World Bank.

Schuster, C. (2026) Using Workforce Records to Evaluate Management and Leadership Trainings in the Civil Service: A Feasibility Assessment, Unpublished Report, London: Cabinet Office.

Uttl, B., White, C., Gonzalez, D. (2017) Meta-analysis of faculty's teaching effectiveness: Student evaluation of teaching ratings and student learning are not related. *Studies in Educational Evaluation, 54*, p. 22-42.

UK Government (2020) *The Magenta Book: Guidance for evaluation*. London: HM Treasury.

Wolf, C., Joye, D., Smith, T., and Fu, Y. (2016a) *The SAGE Handbook of survey Methodology*. SAGE Publications Ltd.

Wolf, C., Joye, D., Smith, T., Fu, Y., and Smyth, J. (2016) Designing questions and questionnaires. In *Designing questions and questionnaires*. SAGE Publications Ltd, pp. 218-235.

Government
Skills

Appendix A: Example Survey Questionnaire: End-of-Programme Senior Practitioner Survey

Professional background

[Note: the pre-programme survey contained a range of additional professional background questions]

Which of the following best fits your grade, pay band or responsibility level?

If your department operates a different grade structure to the one set out below, please choose the grade below that most closely resembles your departmental grade (click here for an overview of departmental grade structures):

SCS1/2 or Director/Deputy Director

G6

G7

HEO

SEO

EO

AA

AO

Other

Don't know / prefer not to respond

How many staff do you line manage? Please only include staff reporting directly to you.

Don't know / prefer not to respond

I am not a line manager

Dropdown, 1-100 or more

Training participation

We would now like to ask you several questions about the training and your participation. There are no right or wrong answers in this section.

How would you rate the following on a scale from 1–10?

(1 = very poor, 10 = excellent)

	1 (very poor)	2	3	4	5	6	7	8	9	10 (excellent)	Don't know / prefer not to respond
The overall quality of the training											
Training facilitator(s)											

As a reminder, the training outcomes for this course are:

- Increase your levels of self-awareness and emotional intelligence
- Show you how to take a coaching style approach to developing your team members and how to create an environment of psychological safety
- Demonstrate how to effectively receive and deliver challenging feedback, both with your colleagues and your stakeholders
- Equip you with an understanding of systems thinking.
- You'll be confident in balancing business-as-usual operations with calculated risk-taking, designed to improve the service your team delivers
- Allow you to use your networks to seek out further perspectives and expertise that can improve your decision-making and lead to increased innovation

To what extent do you agree that the training has met its training outcomes?

Strongly disagree

Disagree

Agree

Strongly agree

Don't know / prefer not to respond / did not participate in session

To what extent do you agree or disagree with the following statements about your learning?
I learned a lot from the…

	Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree	Don't know / prefer not to respond
Training workshops (modules 1–3)						
Final Reflection and Closing Call workshop						
GoGreen Simulation workshop						
Action learning sets						
Training workbooks						
One-on-one coaching session						
Psychometric assessment						

To what extent do you agree or disagree with the following statements about the training?

	Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree	Don't know / prefer not to respond
The different topics covered in the training fitted together well						
The training covered topics in sufficient depth to allow me to apply what I have learned in my line management job						
The training was a useful reminder of a line manager's range of responsibilities						
The training provided me with many new insights into how to be a better line manager						

	Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree	Don't know / prefer not to respond
I found the training platform (LEAP) and resources easy to access and navigate						
This training has given me a clear sense of my areas for development relating to line management						
My job offered me many opportunities to apply what I learned in the training						

How many of the four interactive training workshops (including the simulation) did you attend?
(Please do not count your participation in the launch event)

- 0
- 1
- 2
- 3
- 4
- Don't know / prefer not to respond

Did you participate in the one-on-one coaching session?

- Yes
- No
- Don't know / prefer not to respond

How many of the training's action learning sets did you attend?

0

1

2

3

Don’t know / prefer not to respond

Did you read the relevant workbook for the module before each training workshop?

Yes, for every module

Yes, but only for some of the modules

No, I did not read the workbook before any of the training workshops

Don’t know / prefer not to respond

Approximately what proportion of the tasks in the workbooks did you complete before each of the training workshops?

None of the tasks

Roughly a quarter of the tasks

Roughly half of the tasks

Roughly three quarters of the tasks

All of the tasks

Don’t know / prefer not to respond

Did you spend time with your manager to jointly reflect on your training and to get advice?

Yes

No

Don’t know / prefer not to respond

Have you previously attended a different line management or leadership training programme of a similar or longer length (excluding short induction courses and short line management/leadership courses)?

Please tick all that apply.

- Yes, line management training
- Yes, leadership training
- No, this was my first line management or leadership training programme of a similar or longer length
- Don't know / prefer not to respond

Which of the following best describes your participation in this training programme?

- Voluntary
- Mandated by someone (for example, my organisation, profession, function)
- Don't know / prefer not to respond

Not including the full psychometric assessment you took towards the start of the programme, how many of the growth trackers (the smaller assessments focusing only on one or two areas of the full psychometric assessment) did you take during the programme?

- 0
- 1
- 2
- Don't know / prefer not to respond

Your Understanding of Line Management and Leadership

We would now like to ask you several questions to learn about your understanding of line management and leadership. Please note that some questions have multiple correct responses, while others only have a single correct response. Recall that your individual answers will not be reported or shared with government organisations. Please feel free to answer as you see fit.

Is the following statement true or false? "Empowering teams to analyse and solve their problems is a characteristic of a 'hero' style of management."

True

False

Don’t know

Prefer not to respond

Is the following statement true or false? “Leadership comes from a perspective of internal operational process, focusing on continuous improvement to meet or exceed objectives.”

True

False

Don’t know

Prefer not to respond

Is the following statement true or false? “An inspiring leadership style focuses on developing people for the future.”

True

False

Don’t know

Prefer not to respond

Is the following statement true or false? “An affiliate leadership style creates harmony or emotional bonds within teams.”

True

False

Don’t know

Prefer not to respond

Which of the following shared beliefs tend to contribute to greater perceptions of psychological safety in a team (an environment where individuals feel safe to express themselves, take risks, and share ideas or concerns)?

Please tick all correct option(s). Multiple options possible.

It’s OK to be you no matter your gender, age, social background, neurodiversity or sexual orientation, among others

It's OK to take risks

It's OK to question ideas of senior leadership

It's OK to criticise other team members

It's OK to speak up with questions

It's OK to admit mistakes

None of the above

Don't know

Prefer not to respond

Is the following statement true or false? "Paraphrasing – repeating what a team member has said in your own words and then checking understanding – is an important part of coaching."

True

False

Don't know

Prefer not to respond

Is the following statement true or false? "A light-touch coaching approach is particularly suitable for new learners in the team."

True

False

Don't know

Prefer not to respond

What does the coaching model GROW stand for?

Goal, Review, Opportunity, Will or way forward

Goal, Review, Options, Will or way forward

Goal, Reality, Options, Will or way forward

Goal, Reality, Opportunity, Will or way forward

Goal, Review, Opportunity, Workflow

None of the above

Don't know

Prefer not to respond

Which of the following should you keep in mind when receiving challenging feedback? Please tick all correct option(s).

Multiple options possible.

- Treat the feedback you are receiving as information
- Make your own choices about what you intend to do with the feedback (accept, reject, wholly or partly)
- Let the other person know what you are doing or have done with the feedback
- Avoid arguing, denying or justifying
- Ask for clarification if you do not fully understand the feedback
- Distinguish between the content of the feedback and your reaction to it
- None of the above
- Don't know
- Prefer not to respond

What does a 'VUCA' environment stand for?

- Variability, Unpredictability, Complexity, Ambivalence
- Volatility, Unpredictability, Complexity, Ambiguity
- Volatility, Uncertainty, Complexity, Ambiguity
- Variability, Unknowns, Complexity, Ambivalence
- Volatility, Uncertainty, Complexity, Ambivalence
- None of the above
- Don't know
- Prefer not to respond

Which of the following are typical characteristics of 'adaptive challenges'? Please tick all correct option(s).

Multiple options possible.

- Require solutions by experts or authorities
- Require changes in values, beliefs, relationships and mindsets

Require experiments and overcoming dead ends

Lend themselves to operational (process and procedure) solutions

Require change in numerous places

None of the above

Don't know

Prefer not to respond

In which situations is systems thinking particularly helpful? Please tick all correct option(s).

Multiple options possible.

Disagreement among stakeholders

Complex challenges

Clearly understood solutions

Defined objectives

Interdependencies between different parts of a policy problem

None of the above

Don't know

Prefer not to respond

Attitudes, Experiences and Perceptions

We would now like to ask you several questions about your perceptions of line management and attitudes towards management and the Civil Service more generally. There are no right or wrong answers in this section.

To what extent do you agree with the following statements?

	Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree	Don't know / prefer not to respond
I am satisfied with my job						
When I wake up in the morning, I feel like going to work						

	Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree	Don't know / prefer not to respond
I want to stay working for my organisation for at least the next three years						

To what extent do you agree with the following statements?

	Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree	Don't know / prefer not to respond
I enjoy being a line manager						
I feel motivated by the challenges and responsibilities of my line management role						
I feel a sense of accomplishment from being a line manager						
I care deeply about being a good line manager for the people I manage						
It is important for me to be a good line manager to improve my promotion prospects						

To what extent do you agree with the following statements?

	Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree	Don't know / prefer not to respond
I have the skills and knowledge I need to be an excellent line manager						
I have a strong network in the Civil Service to connect with to seek ideas and external perspectives to improve my line management						
I care deeply about the career development of the staff I supervise						
I adopt a growth mindset at work and try new things without being scared to fail						
My own manager supports me to improve my line management skills						
I am aware of the official Civil Service Line Management Standards						

To what extent do you agree with the following statements?

	Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree	Don't know / prefer not to respond
My team members don't have opportunities to be innovative or creative						
I have a coaching mindset at work, using every available						

	Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree	Don't know / prefer not to respond
opportunity to support the development of my team members						
I consider challenging feedback I receive on my work as an opportunity to help me improve my performance						
I experience a strong sense of togetherness with my team in my everyday work						
I have a clear sense of my own strengths and areas for development as a manager						
I embrace change at work						

How challenging do you find the following in your line management role?

	Not at all challenging	Slightly challenging	Moderately challenging	Challenging	Very challenging	Don't know / prefer not to respond
Trusting team members to complete tasks effectively without close supervision						
Making sure team members have the right capabilities to do their jobs						
Getting the best out of my team						
Handling difficult conversations with						

	Not at all challenging	Slightly challenging	Moderately challenging	Challenging	Very challenging	Don't know / prefer not to respond
team members effectively						
Assessing the big picture of problems I face at work, including interdependencies with other parts of my organisation and other departments						

Line Management Practices

Thinking about the last three months, how frequently have you done each of the following at work?

Please choose the frequency which best applies to you.

	Never	Once or twice in the last three months	Once a month	Two to three times a month	Once a week	Two to three times a week	Once a day	Don't know / prefer not to respond
Seized "in the moment" opportunities to coach team members								
Created dedicated time in the calendar for focused coaching of individual team members								
Observed team members to help them achieve their potential								

	Never	Once or twice in the last three months	Once a month	Two to three times a month	Once a week	Two to three times a week	Once a day	Don't know / prefer not to respond
Asked team members about how they feel								
Listened to challenges team members face and solicited their ideas to address them, without proposing your own solutions								
Identified new activities for team members to take on to help them develop								

How frequently have you done each of the following at work in the last three months?

Please choose the frequency which best applies to you.

	Never	Once or twice in the last three months	Once a month	Two to three times a month	Once a week	Two to three times a week	Once a day	Don't know / prefer not to respond
Micromanaged low performing team members to help them improve								
Communicated to team members that they can and should raise risks without fear of punishment								
Communicated to team members that they can and should challenge the way the team works – including myself – to								

	Never	Once or twice in the last three months	Once a month	Two to three times a month	Once a week	Two to three times a week	Once a day	Don't know / prefer not to respond
come up with improvements, without fear of punishment								
Listened to a team member struggling with a task, without being judgmental								

How frequently have you done each of the following at work in the last three months?

Please choose the frequency which best applies to you.

	Never	Once or twice in the last three months	Once a month	Two to three times a month	Once a week	Two to three times a week	Once a day	Don't know / prefer not to respond
Had challenging conversations with team members without them or I being very emotional								
Took pushback I received when leading change personally								
Let my team members input meaningfully on ideas for my team(s)								
Admitted my own mistakes to team member(s)								
Enabled 'leadership from below', by implementing or supporting solutions to team challenges								

	Never	Once or twice in the last three months	Once a month	Two to three times a month	Once a week	Two to three times a week	Once a day	Don't know / prefer not to respond
proposed by those in lower ranks								

Finally, how frequently have you done each of the following at work in the last three months?

Please choose the frequency which best applies to you.

	Never	Once or twice in the last three months	Once a month	Two to three times a month	Once a week	Two to three times a week	Once a day	Don't know / prefer not to respond
Applied systems thinking to an issue at work, including by mapping out interdependencies between factors affecting your issue								
Took a step back to identify complex challenges at work, which require experimentation, a broader mindset, and relationship shifts								
Brought together diverse stakeholders to make progress on a complex challenge								
Decentred yourself in change processes you are formally in charge of, to enable those closer to the issue to drive the change								
Read articles or books, or watched								

	Never	Once or twice in the last three months	Once a month	Two to three times a month	Once a week	Two to three times a week	Once a day	Don't know / prefer not to respond
videos on management (not as part of other government training) to improve your management skills								

Receive your Personalised Management Diagnostic

Please insert your email address here to receive a personalised management diagnostic based on your survey responses which you will use in your training workshop discussion. Make sure you use the same email address as in the survey you completed at the beginning of the training. You will receive an email within minutes of closing the survey. Please check your spam folder in case the email is not received.

If you have any further comments or suggestions about the training, please do not hesitate to let us know here:

Please click on 'Next Page' below for your response to be counted and to receive your management diagnostic via email.

Appendix B: example interview guide

Programme: Fundamentals Line Manager Training Evaluation

Interview type: Semi-structured participant interview

Interview timing: (post-training, 3 to 6 months)

Duration: 45 to 60 minutes

Opening (~ 5mins)

- Thank participant for their time.
- Explain purpose: understand their experience of Fundamentals training and how it has influenced their management practice.
- Emphasise confidentiality. Responses will be anonymised and reported thematically.
- Confirm consent to record and proceed.

Contextual background (5-10 mins)

- Tell me about your **role** and how long you've been a **line manager.**
- What made you want to participate in Fundamentals? [Probe: **enrolment**]
- What were you hoping to gain from the programme? [Probe: **expectations**]

Training experience and learning outcomes (10-15 mins)

- Thinking back to all the modules, which parts of the training have stayed with you most? Why? [Probe: **significant experiences**]
- Can you describe a moment during the training when something "clicked" or made you think differently about management? [Probe: specific **examples** from programme content on, for example, delegation, feedback, performance management, or inclusion modules]
- Has your **confidence** in any aspect of people management changed since the training? [Probe: explanation. In what way? What contributed to that change?]
- Did you form **connections** with other participants that have continued? [Probe: cohort connections. If yes: How have those relationships been helpful?]

Application and behaviour change (15-20 minutes)

- Can you give me a specific example of a time you **applied** something from Fundamentals in your work? [Probe: what prompted you to use it? What happened?]
- Since the training, has anything changed in how you (pre-identify **expected behaviours**): [Probe: frequency, confidence, approach]
    - Delegate tasks to your team?
    - Give feedback or hold coaching conversations?

    - Handle difficult conversations about performance?
- **Barriers/challenges**: What has made it difficult to use what you learned?
- **Enablers**: What enabled you to implement these changes?

Outcomes and value (5-10 minutes)

- Have you noticed any **changes in your team** since the training? [Probe: team dynamics, engagement, performance, morale]
- If a colleague asked whether they should do Fundamentals, what would you tell them?
- Looking back, what do you see as the **most valuable aspect** of the programme?

Improvement recommendations (5 minutes)

- If you could **change** one thing about Fundamentals to make it more effective, what would it be?
- Was there anything you expected to learn that wasn't covered, or wasn't covered in enough depth?
- [If relevant: **survey findings** to explore] Survey results showed [unexpected pattern]. Why do you think that might be?

Closing (2 minutes)

- Is there anything else about your Fundamentals experience or its impact that we haven't discussed?
- Thank participant and explain next steps (how findings will be used, when report available).

Appendix C: example observation guide

This observation protocol was developed and used in evaluations of Civil Service line management training programmes. It demonstrates a middle-ground approach between highly structured checklists and completely unstructured observation, organising attention thematically while remaining open to emergent patterns.

The protocol derived from programme theories of change and content analysis, directing attention to theoretically salient features of delivery. Observations focused on programme implementation rather than individual performance assessment, maintaining ethical boundaries appropriate for developmental evaluation.

Format: Passive observation (non-participant)

Recording: Field notes during session, expanded immediately post-observation

Duration: Full workshop sessions (typically 3.5-7 hours)

Ethical approach: Observer introduced at session start, voluntary participation emphasised, no individual identifiers recorded

Theme 1: Facilitator Approach

- Opening and framing: How session purpose and structure introduced
- Facilitation style: Directive versus collaborative, energy and engagement
- Content delivery: Balance of presentation versus discussion versus activities
- Responsiveness: Adaptation to participant needs, time management, flexibility

Theme 2: Participant Engagement

- Attention and focus: Cameras on/off, active listening, distraction patterns
- Participation patterns: Who speaks, silence/dominance, group dynamics
- Questions: Clarification versus application versus critical engagement
- Peer interaction: Supportive, competitive, collaborative
- Emotional tone: Enthusiasm, frustration, confidence, vulnerability

Theme 3: Content

- Examples and scenarios: Resonance with participant realities
- Civil Service specificity: Generic versus context-embedded
- Module connections: Coherence across programme elements

Theme 4: Pedagogy

- Activity design: Engagement, learning value, time allocation
- Discussion: Depth, breadth, inclusion

- Skill practice opportunities: Realism, psychological safety, feedback quality
- Theory-practice bridges: How frameworks connected to application
- Resource integration: Resources used, nature of engagement

Theme 5: Logistics & Environment

- Platform functionality: Technical issues, user experience, workarounds
- Time management: Pacing, breaks, session flow
- Virtual/in-person dynamics: Specific advantages or challenges
- Cohort composition: Open versus closed, departmental mix

we have identified any third-party copyright material you will need to obtain permission from the copyright holders concerned.