Part B

Planning and undertaking evaluations

This part of the Magenta Book is written for analysts and interested policy makers and sets out key issues to be considered when developing, planning and undertaking evaluations. It describes the process of planning an evaluation and the collection of supporting data and sets out the requirements of impact and process evaluations in more detail.

Chapter 5: The stages of an evaluation

Chapter 6: Setting out the evaluation framework

Chapter 7: Data collection

Chapter 8: Process evaluation, action research and case studies

Chapter 9: Empirical impact evaluation

Chapter 10: Drawing together and reporting evaluation evidence

5 The stages of an evaluation

Key points

- There are a number of stages in undertaking an evaluation, involving identifying which questions to ask of the evaluation, which type of evaluation is most appropriate to answer them, and when and how the evaluation should be carried out.

- A first important step is planning the evaluation. This will involve specifying the objectives, timeframes, resource requirements, governance arrangements and terms of reference and should consider how evaluation findings will be used, and by whom, since this will affect how an evaluation is undertaken.

- Using the policy "logic model", which explains how the policy is intended to achieve its objectives, is always recommended for any evaluation. This will help to clearly identify the evaluation objectives and research questions which will direct the evaluation approach, and inform the types of data and information that need to be collected.

- The evaluation objectives and research questions should also guide a review of the existing evidence relevant to the research questions.

- While an evaluation will be planned to answer questions of immediate interest, it should also be capable of having a longer-term strategic influence.

Introduction

5.1 Chapter 5 describes the various stages involved in planning, commissioning and undertaking an evaluation. Considering each of these steps before the evaluation is undertaken will help to:

- identify the information requirements for the evaluation;

- ensure an appropriate evaluation approach is adopted;

- identify key dates and milestones; and

- ensure the quality, transparency and policy relevance of the evaluation findings.

5.2 Evaluation planning is an important part of policy design. However, as policy making and evaluation are often iterative; it may be necessary to review some of the evaluation objectives and questions as the project progresses.

5.3 A summary of the steps to be considered in planning and undertaking an evaluation was presented in Chapter 4 and is represented in Table 5.A. The remainder of this chapter discusses each of the steps in more detail.

Table 5.A: Steps involved in planning an evaluation

Defining the policy objectives and intended outcomes	• What is the programme logic or theory about how inputs lead to outputs, outcomes and impacts, in the particular policy context?
Defining the audience for the evaluation	• Who will be the main users of the findings and how will they be engaged?
Identifying the evaluation objectives and research questions	• What do policy makers need to know about what difference the programme made, and/or how it was delivered? • How broad is the scope of the evaluation?
Selecting the evaluation approach	• Is an impact, process or combined evaluation required? • Is an economic evaluation required? • How extensive is the evaluation likely to be? • What level of robustness is required? • Can proportionate steps be taken to increase the potential for good evaluation? • What adjustments to policy implementation might improve evaluation feasibility and still be consistent with overall policy objectives?
Identifying the data requirements	• What data are required? • What is already being collected / available? • What additional data need to be collected? • If the evaluation is assessing impact, at what point in time should the impact be measured? • Who will be responsible for data collection and what processes need to be set up? • What data transfer and data security considerations are there?
Identifying the necessary resources and governance arrangements	• How large scale / high profile is the policy, and what is a proportionate level of resource for the evaluation? • What is the best governance structure to have in place? • What budget is to be used for the evaluation and is this compatible with the evaluation requirements? Has sufficient allowance been built in? • Who will be the project owner, provide analytical support, be on the steering group? • What will the quality assurance processes be?
Conducting the evaluation	• Will the evaluation be externally commissioned or conducted in-house? • Who will be responsible for specification development, tendering, project management and quality assurance? • When does any primary data collection need to take place? • Is piloting or cognitive testing of research instruments required? • When will the evaluation start and end?
Using and disseminating the evaluation findings	• What will the findings be used for, and what decisions will they feed into? • How will the findings be shared and disseminated? • How will findings feed back into the ROAMEF cycle?

The steps involved in planning and undertaking an evaluation

Step 1 - Defining the policy objectives and intended outcomes

5.4 A first step in evaluation planning is to set out the objectives and intended outcomes of the policy, since this provides a clear framework for subsequent steps, and helps identify exactly what the evaluation should assess. This information might already have been developed as part

of a policy appraisal (e.g. the Impact Assessment) or the Rationale and Objectives parts of the ROAMEF cycle.

Developing the Logic Model

5.5 A common method for setting out the policy objectives and intended outcomes is to develop a logic model (also known as "intervention logic" or "programme theory"). A logic model describes the theory, assumptions and evidence underlying the rationale for a policy. It does this by linking the intended outcomes (both short and long-term) with the policy inputs, activities, processes and theoretical assumptions.[1]

Box 5.A: Components of a Logic Model

Kellogg Foundation Logic Model
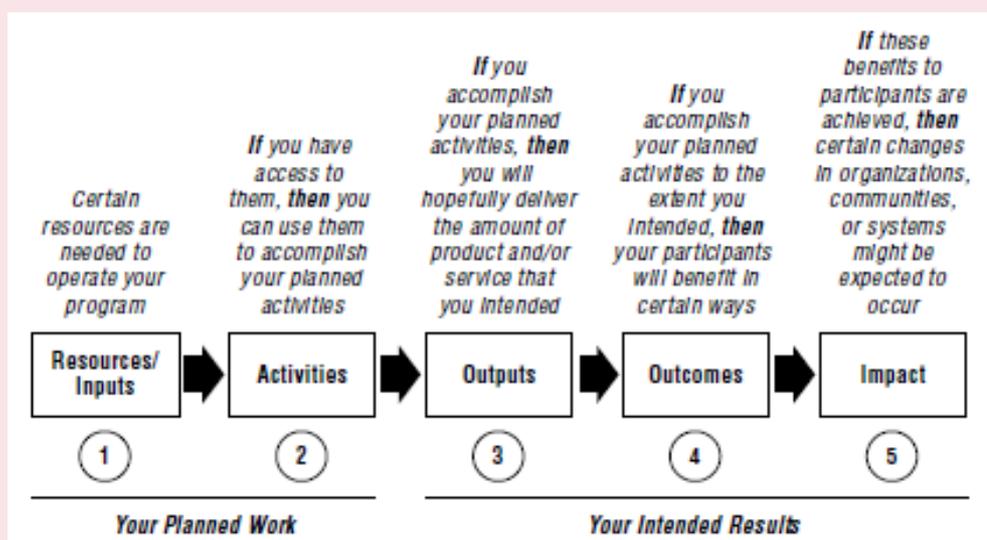
Source: Kellogg Foundation (2004)

5.6 Generally, a logic model will identify the following elements of a policy intervention:

- the issues being addressed and the context within which the policy takes place;

- the inputs, i.e. the resources (money, time, people, skills) being invested;

- the activities which need to be undertaken to achieve the policy objectives;

- the initial outputs of the policy;

- the outcomes (i.e. short and medium-term results);

- the anticipated impacts (i.e. long-term results); and

- the assumptions made about how these elements link together which will enable the programme to successfully progress from one element to the next.

5.7 Setting out the intervention logic model can help to identify clearly the key inputs, and the expected activities, outputs, outcomes and impacts. This is important for a number of reasons, including:

- it can help to guide reviews and collection of existing evidence and data, thereby highlighting areas of deficiency which the evaluation might focus on. Methods for reviewing existing evidence are considered in Chapter 6;

- it can inform the evaluation objectives and development of the research questions;

- it can guide the design of data collection and monitoring processes, so that the right information is available for evaluating the intervention. Data collection is considered in more detail in Chapter 7;

- it can help to identify how the intervention could have unintended consequences, thereby further guiding data collection, the evaluation objectives and the evaluation framework. Unintended consequences are described further in Chapter 6; and

- it provides a transparent assessment framework within which existing evidence and the evaluation results can be combined to provide answers to the evaluation questions.

5.8 There are many ways to produce a logic model (and no necessarily right or wrong approach), but all generally include the elements listed above. Example logic models are described in Chapter 6.

Step 2 – Defining the audience for the evaluation

5.9 To ensure the evaluation provides useful evidence, it is important to consider who the anticipated users of the findings are, and the requirements of policy makers and other stakeholders. These considerations need to be made before the evaluation starts. The findings might be used to:

- support the implementation of the policy;

- inform future decision-making;

- support funding applications;

- improve the ongoing delivery process;

- provide accountability to stakeholders, parliament and the public; and

- contribute to improved knowledge amongst those best able to take advantage of it.

5.10 Thus, when developing the evaluation plan, it is important to understand:

- who the target end-users of the evidence will be. This may include programme managers, policy makers and analysts within the department; other government departments; local authorities and delivery bodies; or key stakeholders including industry bodies, the public, local community groups and other interested parties.

- what are the different expectations for how the results will be used (particularly important for results which may feed back into and affect the ongoing programme delivery) including any expectations on the timing of when the evaluation evidence might feed into decision making;

- what will allow the end users to make most effective use of the evaluation findings. This includes different data requirements, but also presentation of the results, mechanisms for and timing of dissemination. For example, a quantitative cost-

benefit assessment of impacts may be required by HM Treasury, while detailed information about effective delivery may be sought by programme managers responsible for the implementation of the same programme on the ground; and

- how robust the evaluation results need to be to support the uses they are intended for, and what level of scrutiny they will be subject to. A decision to support the potential funding and roll-out of a major government initiative is likely to require a high "burden of proof" and hence an evaluation which meets the highest academic standards. Related to this is whether you expect to use average or marginal effects (see Chapter 10, paragraph 10.15 and Table 10.A for further information). An evaluation which is intended to inform specific and limited changes to the way an existing, local intervention is delivered is unlikely to require the same levels of rigour. However, this might limit the generalisability of the evaluation findings and the extent to which they can be seen to add to the evidence base.

5.11 These considerations are therefore likely to influence the evaluation objectives, research questions and evaluation design. By understanding the range of requirements for the evaluation, the questions can be designed to reflect these and methods can be chosen that generate relevant evidence (Step 3).

Step 3 – Identifying the evaluation objectives and research questions

5.12 The third step in planning an evaluation is to identify the evaluation objectives, and the questions the evaluation will address. The logic model will assist this process by identifying the anticipated inputs, outcomes and impacts. Importantly, the model will also identify theoretical links between inputs and outputs that the evaluation may need to test. When developing the evaluation questions, it is important to assess not only the importance of each question but also how the information will be used. This will help prioritise and determine what is to be evaluated. It will also be necessary to consider what constitutes a proportionate and realistic evaluation given the resources and data available, and what is already known about the policy and its delivery.

5.13 As part of this consideration, when planning the evaluation it is important to decide what the evaluation will add to the existing body of knowledge about what does or does not work. In the case of a new, innovative or pilot policy, this may be fairly obvious. However, in other cases it may be more important for the evaluation to confirm previous results in different contexts, or explore aspects that previous evaluations of similar policies left untouched. In either case, a good understanding of what is already known and the existing evidence base is crucial. If an important question is whether the programme is more effective than similar ones evaluated previously, it will be important to ensure that the evaluation is planned and data collected in such a way as to maximise comparability between the two sets of findings.

5.14 As outlined in Chapter 2, whatever the scope of the evaluation questions, they will normally fall under two broad questions "what difference did the policy make?", or "how was it delivered?" However, it will be necessary to define more specific questions than these; the evaluation questions will be quite specific to the particular policy and logic model. Identifying the evaluation questions is an activity that would normally be undertaken jointly by policy and analytical colleagues. Table 5.B lists a number of issues to consider when developing evaluation questions.

Table 5.B: Issues to consider when developing evaluation questions

What difference did the policy make?	How was the policy delivered?
How will you know if the policy is a success? Which of the outcomes will it be important to assess?	Is it important to understand why the policy does or does not achieve anticipated outcomes?
Do you need to quantify impacts, as well as describe them? How measurable are the various outcomes which might describe the policy's impacts?	Which aspects of the delivery process are innovative or untested?
How complex is the impact pathway/logic model? How important is it to control for confounding factors?	Is it important to learn about uptake, drop-out, attitudes etc.?
What were the impacts for the target group? Do you need data on average or marginal impacts?	What contextual factors might affect delivery (e.g. economic climate, other policy measures, etc.)?
Were there different impacts for different groups?	What process information would be necessary, or useful, for any planned impact evaluation?
How developed is the existing evidence base? Could it enable the scope of the evaluation to be restricted to those areas, impacts or processes where knowledge is most uncertain?	What were the experiences of service users, delivery partners, staff and organisations?
How should the costs and benefits of the policy be assessed? How do the outcomes contribute to social wellbeing, and how do they generate costs?	How complete are current data collection processes? Are the issues to be considered likely to need tailored data collection?
What longer term or wider knock-on effects should be considered? How will you know whether there were any unintended effects?	How was the policy delivered?

Step 4 - Selecting the evaluation approach

5.15 There are a variety of approaches to evaluation, which differ in a number of respects. These include the analytical techniques they adopt, the types of data they use, and the nature of the results they generate. Box 5.B provides a brief description of some of these broad approaches. These categories are not necessarily distinct; however each can comprise a number of different approaches.

Box 5.B: Types of evaluation

Process evaluation

Process evaluations can use a variety of qualitative and quantitative techniques to explore how a policy was implemented describing the actual processes employed, often with assessments of the effectiveness from individuals involved or affected by the policy implementation. Further discussion appears in Chapter 8.

Empirical impact evaluation

Empirical impact evaluations use quantitative data to test whether a policy was associated with any significant changes in outcomes of interest. Various approaches are available which differ in their ability to control for other factors which might also affect those outcomes (the counterfactual, either directly measured or imputed) and hence in the confidence it is possible to place in the results. Empirical impact evaluation is discussed further in Chapter 9.

Economic evaluation

Economic evaluation involves calculating the economic costs associated with a policy, and translating its estimated impacts into economic terms to provide a cost-benefit analysis. (When only a costing exercise is undertaken, the result is a cost-effectiveness analysis.) Economic evaluations will often make use of existing evidence and assumptions to facilitate the translation of inputs and actual measured outcomes into economic measures, making them akin to theory-based evaluations (see below). The HM Treasury Green Book provides detailed guidance on economic evaluation and cost-benefit analysis.

Theory-based evaluation

Theory-based evaluation approaches involve understanding, systematically testing and refining the assumed connection (i.e. the theory) between an intervention and the anticipated impacts. These connections can be explored using a wide range of research methods (both qualitative and quantitative), including those used in empirical impact evaluation. More information is provided in Chapter 6.

Meta-evaluation and meta-analysis

Meta-evaluations (covered in more detail in Chapter 6) can use quantitative or qualitative techniques to bring together a number of related evaluations to derive an overview or summary conclusion from their results.

Simulation modelling

Simulation modelling is one way in which the results of different evaluations of separate parts of the impact pathway or logic of an intervention can be combined and requires that the evidence relating to the different links in the logic model are expressed in quantitative terms (e.g. effect sizes). Chapter 6 provides more information.

5.16 The choice of evaluation approach will depend on a number of factors, some of which are considered in Table 5.C. The exact evaluation approach will generally be developed by analytical colleagues, and/or recommended by an evaluation contractor (for externally commissioned evaluations) or other evaluation expert. However, having a clear idea about the required type of evaluation at the planning stage will help inform its design and ensure this meets the evaluation requirements. This will greatly aid decisions about the scope and scale of the evaluation, development of the specification, and the external expertise required.

5.17 There are therefore a wide range of evaluation approaches which will be more or less suitable to the specific evaluation questions and context. Process evaluation is discussed in more detail in Chapter 8 and experimental and quasi-experimental impact evaluation approaches are discussed in Chapter 9. Systematic review, meta-evaluation, theory-based approaches and simulation modelling are discussed in Chapter 6.

Box 5.C: Issues affecting the choice of evaluation approach

Evaluation objectives and research questions

The overall objectives of the evaluation and the specific research questions it needs to answer are important factors in deciding which evaluation approach(es) to use and should be developed from the logic model. General research questions which are not overly specific to the intervention in question might be answerable via a qualitative review (or more formal analysis) of the existing literature. Questions which are more specific to the intervention will involve one of the other evaluation types listed in Box 5.B. Questions relating to the wider or ultimate objectives of an intervention will generally require some form of impact evaluation – possibly as part of a theory-based evaluation approach if the associated impact pathways are very extended or complex. Questions relating to detailed aspects of the workings of the policy will generally imply some form of process evaluation (although a combined impact evaluation might be warranted if more definitive answers about effectiveness are required).

Complexity of the logic model and importance of confounding factors

Where the logic model is particularly complex, restricting the scope of the evaluation to consider shorter, simpler "links" in the logic chain can increase the ability of process evaluations to provide good evaluation evidence. However, if significant confounding factors remain, a robust impact evaluation with suitable controls might be necessary to generate reliable findings. The feasibility of this might depend on data availability (for quasi-experimental approaches) and time and resources (for approaches needing dedicated data collection). Detailed evaluation of changes in very complex systems (especially those with a significant geographical component) might only be possible through theory-based evaluation or simulation modelling.

Availability and reliability of existing evidence

Large amounts of strong existing evidence increase the relevance of review based methodologies, facilitate greater use of simulation models, and enable evaluations to be simplified to focus more closely on those specific questions which the current evidence base leaves unanswered.

Existing data sources and measurability of outcomes

If there is already a wide range of good quality data sources covering outcomes of interest, the feasibility of undertaking robust impact evaluations (sometimes to relatively short timescales) is greatly increased. Outcomes which are difficult to measure require either dedicated data collection (e.g. through surveys) or a way of estimating them from changes in intermediate indicators. The former implies a more resource- and time-intensive study, as does a lack of existing data (which might be the case particularly when the focus of the evaluation is the specifics of a very localised intervention). The latter might be addressed through a simulation model, subject to existing data availability.

Time and resource availability

In most cases, process evaluations (including action research and case studies) will require a formal commission and a dedicated research team, often externally contracted. This can imply a considerable time and resource commitment. Impact evaluations requiring specific data collection and outcome measurement can similarly involve heavy resource commitment and long project durations. Impact evaluations which are able to use existing datasets can provide rigorous results in relatively short timescales but this same reliance on existing data can restrict the questions they can attempt to answer and, in some cases, the ability to confidently attribute the impacts to the intervention. Simulation models can also sometimes be undertaken relatively quickly but this depends on a range of assumptions being made to limit their scope.

Empirical impact evaluation issues

The two principal strengths of empirical impact evaluation approaches are that they can isolate the effect of an intervention from the possible multitude of factors which might have an influence on the outcome of interest; and in this way, they can provide a rigorous test of whether the intervention has an effect or not. However, these strengths can come at a cost. That is that the approaches are often less able than other approaches to explain exactly why any difference occurred (or not), or how it varied across circumstances.[2] Much of this can (and should) be overcome by using a mixed design, whereby process and impact evaluations complement each other, and the process evaluation can help to explain the impact evaluation findings.

In other cases based on statistical regression analysis the relationship between the intervention and the outcome of interest might be so complex that the evaluation will only be able to say whether the intervention had an effect, not what aspects of it, how or why. Some "procedural" explanation might be possible, but only if the scope of the evaluation is restricted to simpler relationships, for instance, between the intervention and some intermediate outcome rather than the ultimate objective of the intervention (e.g. the impact of the intervention on the take up of training, rather than the impact on employment and wages).

Step 5 – Identifying the data requirements

5.18 A good evaluation relies on good quality data. The evaluation questions will determine what data need to be collected, and when. This may be new data but will often also include monitoring data, that is, information collected and used as part of the ongoing policy delivery, describing the principal policy inputs and outputs (e.g. training sessions provided and completed). (For more information on planning and collecting monitoring data, see Chapter 7.)

5.19 Data requirements may also include data collected specifically for the evaluation through specially commissioned surveys and interviews with participants and frontline workers, and covering the details of the way the policy has been implemented. Evaluations of large scale policies might well also use data which already exist or are being collected for other purposes,

---

[2] Regression-based analysis of data obtained from randomised control trials might be able to provide some explanation of how an observed impact varies across subjects, but is still limited in its explanatory power, and subject to the other weaknesses of the counterfactual impact evaluation approach.

for instance relating to local and regional economic conditions and performance (e.g. sectoral unemployment rates).

5.20 The specific data required for an evaluation will relate to the inputs, outputs, outcomes and impacts of the policy, and when these are expected to manifest. These will have been identified in the first step of planning the evaluation. Data collection processes will reflect the nature of the outcomes in question – outcomes which are unusual (e.g. impacts on individual economic wellbeing) or very specific to the intervention are likely to require special measurement through, for instance, dedicated surveys. Evaluation data may also relate to information about how the various elements of the policy are linked together, the actual delivery process and timescales.

5.21 Data collection will often need to commence before the policy is actually implemented, in order to ensure that the situation before the policy can be captured (also known as the "baseline"). Planning for data collection will obviously need to take place before this and so should be considered as early as possible. The timing of the data collection also needs to be considered carefully – eventual impacts of a policy may take many years to materialise, which are likely to be too distant to be collected as part of an evaluation project. In such cases it may be important to build in collection of data related to intermediate or proxy outcomes which can be used to measure impact in a shorter timeframe. These outcomes might then be "translated" into final outcome measures using the logic model framework.

Step 6 – Identifying the necessary resources and governance arrangements

Securing Resources

5.22 As set out in Part A of the Magenta Book, an evaluation should be proportionate to the scale, risk and profile of the policy, and the extent of the existing evidence base related to the effects of the policy and/or delivery process. Judgements need to be made about the scale and form of evaluation that is required for a particular policy, including whether it should be commissioned externally or conducted (either partly or wholly) in-house. Having a clear idea about the available resources for the evaluation will also influence selection of the most appropriate evaluation approach.

5.23 In some circumstances, it may be useful to undertake a scoping or feasibility study to support this decision making process and assess whether particular evaluation methods are possible. This can foster greater understanding of what can and cannot be evaluated, and therefore what level of investment is required, and can support the development of an appropriate evaluation design. It is also important to consider whether an evaluation requires external evaluators in order to ensure objectivity and transparency. Chapter 4 provides more detail on the factors that should be taken into account when deciding how much resources should be dedicated to an evaluation.

5.24 Evaluations, whether conducted internally or commissioned to an external contractor, will often require significant input to ensure they are designed and delivered successfully. For larger evaluations involving dedicated data collection, this will generally require an appropriate internal project manager with the relevant skills to oversee the evaluation, a senior responsible owner (SRO) or project director, and a steering group to govern the evaluation (Table 5.C).

5.25 The level of input required of different members of the project team will be greatest at key points (in particular, the design, commissioning and reporting stage), but there will be an ongoing resource requirement even if the project is externally commissioned and this should not be underestimated.

Table 5.C: Examples of typical evaluation governance responsibilities

Internal project manager	Senior Responsible Owner/ Project Director	Steering group
Drafting a project specification	Ensuring appropriate resources are committed to the evaluation	Ensuring delivery of a high quality and policy-relevant evaluation
Obtaining any necessary data security clearance	Ensuring the information necessary for the evaluation is collected and made available to the evaluators	Providing advice on how to proceed in the event that circumstances change
Commissioning (if appropriate)	Ensuring the relevant policy makers and analysts are prepared to engage in setting the evaluation questions, contribute to the design of the evaluation methods and interpretation of its results, and take custody of its findings and conclusions	Facilitating the work of external evaluators
Day-to-day management, including management of risks		Providing access to information and contacts
Ensuring the evaluation stays on track, meets its objectives, is on time and is delivered within budget		Quality assuring the research design and suggesting evaluation questions, methods and research tools
Advising any contractors and reacting to issues that develop		Assisting in the analysis and interpretation of the emerging evidence
Quality assuring or arranging for quality assurance of intermediate and final products (e.g. project design, research instruments, final reports and presentations)		
Ensuring the findings are fed back to the relevant audience		

Step 7 – Conducting the Evaluation

5.26 Once the policy objectives, intended outcomes, evaluation approach, and data and governance requirements have been established there are a large number of ongoing project management decisions and tasks to be undertaken to ensure that the evaluation is delivered effectively. Typical considerations might include (see also Table 5.C above):

- deciding whether the evaluation be externally commissioned or conducted in-house;

- developing a specification for the evaluation – see below;

- tendering the evaluation, including agreeing the nature and price of any contract with an external provider;

- providing day to day project management support;

- advising any contractors and reacting to issues that develop;

- identifying project risks and mitigating actions;

- budget management;

- agreeing when any primary data collection needs to take place;

- ensuring appropriate quality standards are met;

- deciding whether or not piloting or cognitive testing of research instruments is required;

- agreeing input to and overseeing quality assurance of evaluation processes and products, for example field work activity, research instruments, data set preparation (e.g. cleaning and weighting), data analysis, presentations or reports;

- ensuring any baseline data is collected;

- agreeing when the evaluation will start;

- agreeing and ensuring delivery against key milestones;

- reporting back to stakeholders and steering groups; and

- agreeing when the evaluation will end.

Defining the Project Specification

5.27 As part of the evaluation planning process a project specification (or terms of reference) for the evaluation should be developed. This should cover the scope and objectives of the evaluation, as well as how it will be conducted, governed and managed, and the delivery of the required outputs.

5.28 The exact content will need to be determined by the evaluation commissioner and/or project manager, and will also need to follow existing departmental procedures and guidance for commissioning and managing research and evaluation. However, it is suggested that the following should be included:

- the background, rationale and objectives of the policy to be evaluated, its target recipients, delivery method and intended outcomes;

- the extent of the existing evidence base related to the policy;

- the evaluation objectives and research questions;

- the audience and intended use of the evaluation;

- the available information, for example monitoring data collection processes already set up;

- the possible evaluation approach, research design and methods;

- the required capabilities, skills and experience of the proposed evaluation and team;

- the required evaluation outputs (including datasets) and the milestones to be met;

- data archiving requirements;

- the indicative budget (if being commissioned externally and consistent with departmental, or other, procurement protocols); and

- the evaluation timetable.

Step 8 – Using and disseminating the evaluation findings

5.29 At the time of planning an evaluation it is a good idea to give some thought to how the findings will be used and disseminated. Different departments will have their own protocols and local arrangements which should be followed.

5.30 As well as taking into account the publication process it is important to consider how findings will be presented and to whom. For example whether there will be one long report, an executive summary, a technical report, and/ or presentations. If you are externally-commissioning the evaluation you will need to specify the format of the report and any presentations at the time of commissioning.

5.31 It is also important to consider how findings will be fed back into the policy process to influence future decision making. In summary, it is important to properly plan an evaluation in advance in order to ensure that it meets the required objectives, collects robust evidence which can answer the specific policy questions, and the findings are disseminated and accessible to the relevant audiences. The remaining chapters in Part B describe evaluation design in more detail.

6 Setting out the evaluation framework

Key points

- The evaluation of an intervention requires a framework within which the evaluation can be designed, data analysed and results interpreted. This framework will generally be based on the intervention's logic model and decisions made about the evaluation objectives.

- Developing the logic model enables the assumptions, processes, impacts and outcomes (both intended and unintended) of the intervention to be identified and articulated, which in turn helps to identify the evidence required to answer the evaluation questions.

- Reviewing existing evidence relating to the broad evaluation questions is important for enabling the objectives of any new evaluation research to be identified and refined. Systematic review, rapid evidence assessment and meta-evaluation are approaches to assessing existing evidence.

- Many evaluations of complex interventions or impact pathways will require a theory-based evaluation framework which seeks to triangulate evidence from multiple sources to test and refine the assumptions made in the logic model. Within this framework the evaluation could draw on evidence gathered through process evaluations and counterfactual impact evaluations as well as using analytical techniques, such as simulation modelling.

- Simulation models can be used to combine existing and new evidence to answer the evaluation questions, but can be subject to some uncertainty due to the need to make assumptions about how the different pieces of evidence are related.

Introduction

6.1 Establishing a framework for the evaluation provides a consistent and systematic means to designing the evaluation, collating and analysing the existing evidence and the new data created, and generating and interpreting the results. It can be used to understand what existing evidence tells us and to identify those gaps in the evidence base which the evaluation should focus on. As suggested in Chapter 5, the evaluation framework is most likely to be based on some form of logic model. This chapter provides more detail on logic models, and how they can be used and developed into a theory-based approach. It also considers some of the techniques which can be used to review and evaluate existing evidence.

6.2 Even if there was a significant body of evidence and experience on which to draw, the rationale for an intervention will have been based to some extent, and in certain aspects, on assumptions about how the inputs will cause the intended outcomes and impacts and what other contextual factors will influence this. These assumptions and the evidence on which they were based can be set out formally in a structured "logic model", which can provide the framework within which the impacts of the intervention can be evaluated and (if appropriate) quantified.

6.3 A logic model describes the causal pathways underlying the rationale for a policy. It does this by linking the intended outcomes (both short and long-term) with the policy inputs, activities, processes and theoretical assumptions. Box 6.A presents the simple Kellogg Foundation logic model, and provides definitions (with examples) of its various components.

Box 6.A: Logic models and the terms they use

A logic model describes the theory, assumptions and evidence underlying the rationale for the programme . . . "*it links outcomes (both short and long-term) with programme activities/processes and the theoretical assumptions/principles of the programme*."
Source: WK Kellogg Foundation (2004)[1]

Kellogg Foundation Logic Model
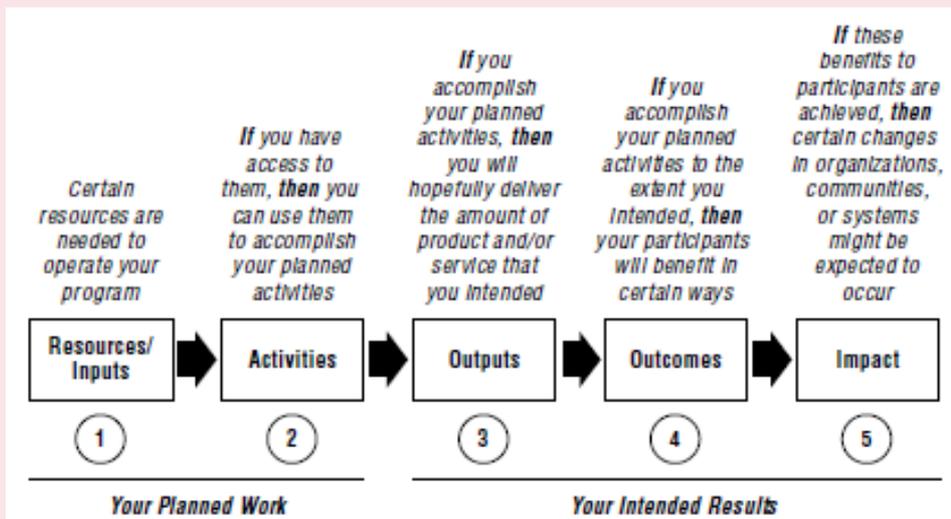
Term	Definition	Example
Inputs	Public sector resources required to achieve the policy objectives	Resources used to deliver the policy
Activities	What is delivered on behalf of the public sector to the recipient	Provision of seminars, training events, consultations etc.
Outputs	What the recipient does with the resources, advice/ training received, or intervention relevant to them	The number of completed training courses
Intermediate outcomes	The intermediate outcomes of the policy produced by the recipient	Jobs created, turnover, reduced costs or training opportunities provided
Impacts	Wider economic and social outcomes	The change in personal incomes and, ultimately, wellbeing

---

6.4 Developing the logic model can be done as a desk exercise, based on a review of policy documentation such as the Impact Assessment, business case or project initiation documents. It may also be developed using previous evaluations and evidence about particular aspects of the inputs, activities, outputs, outcomes and impacts. It might draw on relevant theoretical and empirical frameworks describing the (different links of the) model's impact pathways. Examples could include economic models of individual market behaviour, bio-physical models of the

---

[1] *Logic Model Development Guide,* WK Kellogg Foundation, 2004

impact of air pollution on the environment, and inter-disciplinary models of how changes in health status affect social and psychological wellbeing. Some examples of logic models are provided in Box 6.B which demonstrates that they can be formulated in different ways, albeit around the same basic structure.

Theory-based evaluation

6.5 The examples presented in Box 6.B show that, with different levels of detail, logic models seek to explain how the linkages work between the stages of the logic model as well as simply stating what they are. This is an important feature of an effective logic model, namely that it is a representation of the causal theory underlying the impact and any associated intervention – i.e. the understanding about why something occurs and how an intervention might work. Logic models such as this are therefore an important component of the general class of evaluation approaches called "theory-based evaluation".

Box 6.B: Examples of logic models

Logic Model for the Price Marking Order 2004
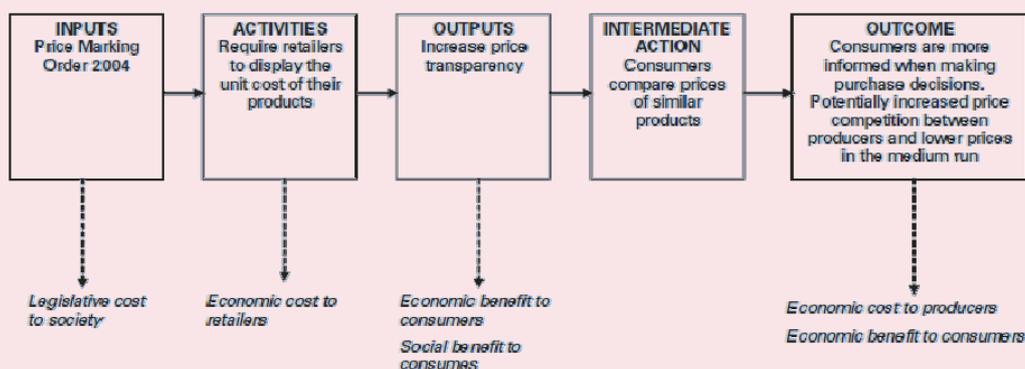
Logic model for the Sector Skills Councils
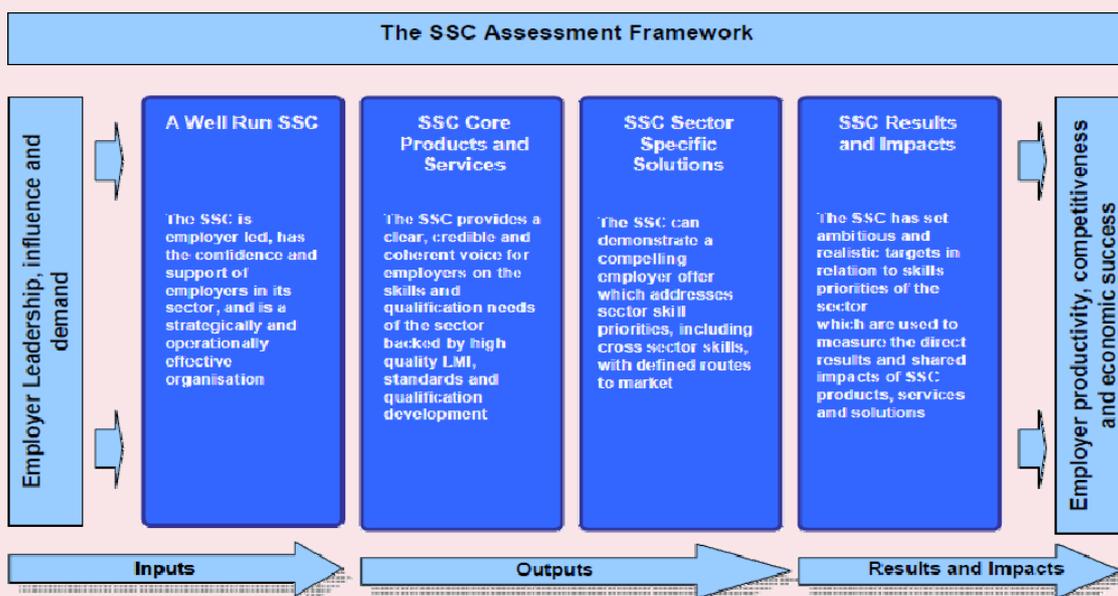
Impact pathway for the health costs of air pollution
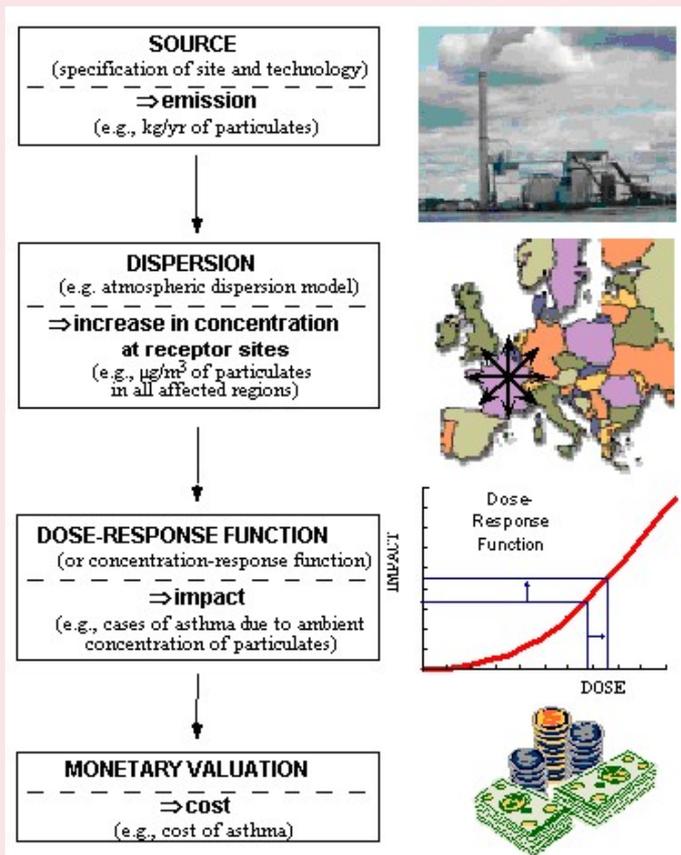
Logic model for Local Area Agreements and Local Strategic Partnerships
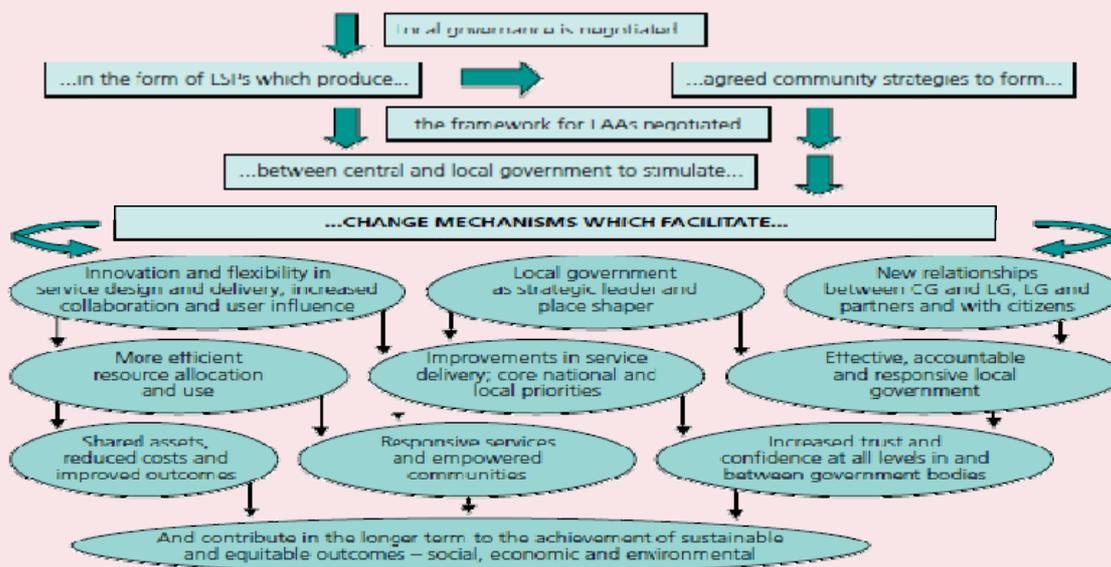
Source: Department for Trade and Industry (2008), UK Commission for Employment and Skills (2009), Department for Communities and Local Government (2008), Externe[2]

---

[2] *The Impact of Regulation: A pilot study of the incremental costs and benefits of consumer and competition regulations,* Department of Trade and Industry, 2006,; *SSC Performance Management Handbook,* UK Commission for Employment and Skills, 2009; *Long Term Evaluation of Local Area Agreements and Local Strategic Partnerships: Developing a 'Theory of Change',* Department for Communities and Local Government, 2008; ExternE – Externalities of Energy: A Research Project of the European Commission (http://www.externe.info/)

6.6 Theory-based evaluation approaches provide an overarching framework for understanding, systematically testing and refining the assumed connections (i.e. the theory) between an intervention and the anticipated impacts.

6.7 The focus of theory-based evaluations is not only on understanding whether a policy has worked, but why, and under what conditions a change has been observed. Theory-based evaluation will therefore generally seek to identify each of the various elements in the underlying logic model, and examine the links between each element. This process is intended:

- to identify clearly the key inputs, and the expected activities, outputs, outcomes and impacts;

- to articulate how inputs are expected to lead to outputs, outcomes and impacts, and the links and processes in place. These are sometimes called "impact pathways";

- to identify the assumptions about how the policy will be delivered, and any additional factors which need to be in place for the policy to succeed;

- to provide a transparent assessment framework for the evaluation to inform the scope, purpose and data requirements of the evaluation; and

- to inform the evaluation objectives and development of the research questions.

6.8 Evaluations of policy issues and interventions within social settings will generally be based on theories of how individuals, groups, organisations and institutions will respond to the intervention given the context in which it is implemented. Three types of theory-based evaluation approach are commonly used in the evaluation of government social policy. Two of these – Theory of Change and Realist (also known as Realistic) Evaluation – are described in Box 6.C below. Box 6.D provides an example of a theory-based evaluation.

Box 6.C: Theory of Change and Realist Evaluation

Theory of Change Evaluation

Theory of Change evaluation is a systematic and cumulative study of the links between activities, outcomes, and context of a policy intervention. It involves the specification of an explicit theory of "how" and "why" a policy might cause an effect which is used to guide the evaluation. It does this by investigating the causal relationships between context-input-output-outcomes-impact in order to understand the combination of factors that has led to the intended or unintended outcomes and impacts. Theory of Change therefore normally develops and tests, the implementation theory of the policy and allows this to be modified or refined through the evaluation process. A range of research methods, often both quantitative and qualitative, can be used in order to gather data that contribute to this task. The evaluation often leads to a map showing which factors at which levels have combined to produce the observed outcomes, building on the logic model for the policy.

Realist Evaluation

Whilst Theory of Change tests implementation theory, Realist Evaluation seeks to identify those – often psychological – triggers that change human behaviour as a result of an intervention, taking into account the context within which the intervention sits. Realist Evaluation typically asks: "what works, for whom, under what circumstances?" It begins by developing a set of hypotheses (or theories) on those factors or processes that explain why an intervention has had a particular result (called a mechanism), and what effect the context of an intervention has on these mechanisms. A mechanism can be defined as capturing "people's reasoning and their choices. They describe how people react when faced with a policy measure".

Source: DfT (2010); Befani B et al (2007)[3]

Box 6.D: Using Theory of Change to evaluate investment in cycling

The Department for Transport has commissioned an evaluation of investment in initiatives aimed to increase cycling in 12 areas across England. This employs a Theory of Change evaluation approach, to assess the impacts of investment and also to learn about what works, how, and why in enabling behaviour change. Each local area has developed investment strategies in response to local need and contextual circumstances. The holistic nature of the approach enables it to test the complex causal relationships involved in changing travel behaviour.

The evaluation is seeking to triangulate evidence to strengthen conclusions about the impacts which can be attributed to the investment programme. The evaluation draws on a quantitative assessment of behaviour and attitudinal change, objective monitoring of cycling trends, analysis of cycling behaviours in comparator areas, qualitative insight into the motivators and barriers to behaviour change, an understanding of the effectiveness of different types of initiatives in overcoming these barriers, an assessment of the role of wider national and local contextual factors and an analysis of the design and delivery processes to identify the barriers and enablers to successful implementation.

Source: Department for Transport (2011)[4]

6.9 The third commonly used theoretical framework for modelling the effects of social policy is the economic model. This model emphasises the role of choices and incentives in driving behaviour of individuals and organisations. Colleagues from the Government Economic Service can provide assistance in developing logic models and evaluation frameworks which incorporate economic principles. The Treasury Green Book can also provide guidance on developing economic evaluations.

6.10 Theory-based evaluations seek to systematically test and refine the underlying logic model. As Box 6.B demonstrated, these logic models are often highly complex, and evaluations based on them will often need to consider large numbers of relationships, and significant quantities of

---

[3] *Guidance for transport impact evaluations: choosing an evaluation approach to achieve better attribution*, The Tavistock Institute for the Department for Transport, 2010, http://www.dft.gov.uk/; *Realistic Evaluation and QCA: Conceptual Parallels and an Empirical Application*, Befani, B et al, 2007, Evaluation, Vol. 13 No 2, p. 178

[4] *Evaluation of the Cycling City and Towns Programme*, AECOM, the Centre for Transport and Society, and the Tavistock Institute for the Department for Transport, January 2011, http://www.dft.gov.uk/

diverse existing evidence and data, including evidence gathered through (existing and new) process evaluations and counterfactual impact evaluations.[5]

6.11 Theory-based evaluation approaches provide the overarching conceptual framework within which specific evaluation studies can be designed and evidence structured to answer the policy questions which are being posed. They are therefore complementary, rather than an alternative, to primary process and impact evaluation studies, which provide new data and evidence which can then be incorporated into the evaluation framework as appropriate. One practical way in which this can be done with quantitative evidence and data is through simulation models (see below).

Assessing wider effects and unintended consequences

6.12 A policy might have wider impacts, such as knock-on or multiplier effects[6] in the local economy.[7] Developing the logic model of the intervention and considering the various stages in which it is intended to operate provides an opportunity to consider the wider or additional effects of the activity. These can then be incorporated into the evaluation as appropriate.

6.13 There might also be effects which are recognised as possible but not definite outcomes of the policy, and which evaluations will also need to look for. They could be harmful or beneficial and might be generated amongst those directly targeted by an intervention or more widely for others indirectly affected by the intervention. Table 6.A presents examples of potential unintended effects.

Table 6.A: Examples of potential unintended effects

Effect	Definition	Example
Displacement	Positive outcomes promoted by government policy are offset by a negative outcome of the same policy elsewhere.	The displacement of crime from one area, where a crime reduction policy is being implemented, to a bordering area.
Substitution	The effects of an intervention on a particular individual, group or area are only realised at the expense of other individuals, groups or areas.	An employer appointing a jobless person from a government scheme, rather than a standard applicant, in order to secure a recruitment subsidy.
Leakage	The policy benefits others outside the target area or group.	Jobs generated in a target area are taken by those who live outside it.
Deadweight	The policy supports outcomes which would have occurred anyway.	An employer receives a subsidy to take on workers who were going to be appointed anyway.

6.14 A policy might also result in other effects that are completely unanticipated, generally termed "unintended consequences". These often result from perverse incentives which are established as a result of interaction between the way the policy works and existing processes. Box 6.E provides examples of unintended consequences and sources of further information.

---

[5] For more information on theory-based evaluations, see *Guidance for transport impact evaluations: choosing an evaluation approach to achieve better attribution*, The Tavistock Institute for the Department for Transport, 2010, http://www.dft.gov.uk/

[6] Further economic activity (jobs, expenditure or income) associated with additional local income and local supplier purchases as a result of the intervention.

[7] For more information see: *Additionality Guide,* English Partnerships http://www.thesroinetwork.org/component/option,com_docman/task,doc_view/gid,30/, *Research to improve the assessment of additionality,* Department for Business, Innovation and Skills, October 2009, http://www.bis.gov.uk/; *wider economic benefits in transport appraisal, Department for Transport* http://www.dft.gov.uk/; and http://www.hm-treasury.gov.uk/green_book_guidance_regeneration.htm

<div style="border: 2px solid #c00; padding: 20px; background-color: #fce8ec;">

Box 6.E: Examples of unintended consequences

The effects of licence plate rationing in Mexico
The most extensive and objective documentation of the long-term impacts of licence plate rationing was found for Mexico City. It was found that there was no sustained improvement in air quality, no increase in subway ridership, and worsening air quality on weekends and other times outside of the rationing scheme.

Modal shift in travel patterns was primarily to taxis and small buses rather than to subways, offsetting any improvements likely to be achieved by reductions in car travel. Demand for petrol went up after two months of implementation, and Mexico City became a net importer rather than net exporter of used vehicles from the rest of the country. The inference was drawn that residents evaded the restrictions by becoming multi-vehicle households (with variably coded licence plates) and acquiring older (and less fuel efficient and more polluting) vehicles from the countryside.

Source: Cambridge Systematics (2007)[8]

The impact of funding incentives on fire prevention
The 2002 Bain Review pointed out a perverse funding incentive that saw the fire authorities dealing with the most fires get the most money. This, along with the tiny amount of funding allocated to fire safety work, did little to raise the profile of community fire safety at a local level. In parallel with Bain's report, work was being done to change the funding model, and from April 2003 the number of fires, false alarms and special calls was removed from the formula. This abolished the perverse incentive that had discouraged a greater focus on fire prevention.

Source: Department for Communities and Local Government (2008)[9]

Further examples are available from: *Additionality Guide,* English Partnerships, (http://www.thesroinetwork.org/component/option,com_docman/task,doc_view/gid,30/), *Research to improve the assessment of additionality,* Department for Business, Innovation and Skills, 2009 (http://www.bis.gov.uk/) and *Wider economic benefits in transport appraisal, Department for Transport*, (http://www.dft.gov.uk/)

</div>

Reviewing the existing evidence

6.15 The first stage in populating the evaluation framework should be to establish what is already known about the intervention to be evaluated or what can readily be learned about it. This first stage is important for at least four reasons:

- it may be that there is already sufficient evidence on the likely effectiveness of an intervention so that further primary evaluation is unnecessary;

- it is more likely that the existing evidence may be ambiguous, inconclusive, or of uncertain quality indicating that further evaluation is necessary and that specific aspects of the policy intervention in question need addressing;

- any single evaluative study may illuminate only one part of a policy issue, implying that it might be appropriate to focus an evaluation on specific aspects of the evidence base where existing information is lacking; and

---

[8] *Congestion Mitigation Commission Technical Analysis: License Plate Rationing Evaluation for the New York City Economic Development Corporation and New York City Department of Transportation,* Cambridge Systematics, 2007
[9] *Safer Houses: Celebrating 20 years of fire prevention in the home*, Department for Communities and Local Government, 2008

- existing findings may be sample, time or context specific. This will make it difficult to establish the generalisability and transferability of findings from the existing research evidence which, in turn, will influence what requires evaluating.

Systematic review

6.16 Establishing what is already known about a policy intervention presents a major challenge for knowledge management. In the first place, the sheer amount of potential research evidence makes it almost impossible to keep abreast of the research literature in any one area. Second, research and information is not of equal value. Some way of differentiating between high and lower quality studies, as well as relevant and irrelevant evidence, is required.

6.17 Systematic review is a tool which can be characterised by:

- a clearly stated set of objectives with pre-defined eligibility criteria for studies;

- an explicit, reproducible methodology;

- a systematic search that attempts to identify all studies that meet the eligibility criteria;

- a formal assessment of the validity of the findings of the included studies; and

- a systematic presentation, and synthesis, of the characteristics and findings of the included studies.

6.18 Systematic reviews therefore differ from other literature reviews by following an explicit protocol for identifying and assessing relevant studies. For instance, the protocol might specify what reference databases were searched, what search terms were used, and what criteria were used to filter studies and select those for detailed review. In general, the review of those studies which are selected will be qualitative (although systematic review can be combined with other evaluation techniques, such as meta-analysis). The basic principles of systematic review are set out in Box 6.F.

6.19 The Campbell Collaboration (http://www.campbellcollaboration.org) provides extensive guidance on undertaking systematic reviews. The Centre for Evidence-Informed Policy and Practice in Education (the EPPI-Centre) (http://eppi.ioe.ac.uk/) at the Institute of Education undertakes and commissions systematic reviews in education, and is developing methods for undertaking systematic reviews of social science and public policy research. The Economic and Social Research Council (ESRC) has established an Evidence Network (http://www.evidencenetwork.org).

Defining an answerable question

A systematic review should address a question that clearly specifies the interventions, factors or processes of interest, the population and/or sub-groups in question, the outcomes that are of interest and the context in which they are set. The question needs to distinguish whether the interest is in the outcome of an intervention or in the implementation of a policy.

An example of an answerable question about a policy intervention might be: What is the effect of a personal adviser service (intervention) in terms of retaining (outcome 1) and advancing (outcome 2) lone parents (population) in the UK workforce (context)?

An example of an answerable question about implementation might be: What are the barriers to (factor/process 1) and facilitators of (factor/process 2) getting lone parents (population) to participate (outcome 1) and advance (outcome 2) in the UK workforce (context)?

Systematic searching for studies

Systematic reviews differ from traditional reviews in the comprehensiveness and procedural formality of searching for all of the available research evidence. It counters problems of selection bias that come from only identifying studies that are readily accessible, or that are only published and indexed in major databases. It also helps reduce publication bias, which comes from the tendency for there to be a higher probability that studies that report positive (or in some cases negative) results are published. Systematic searching involves electronic sources, print sources, and the "grey" literature.

Methods of systematic searching

There are at least two methods of systematically searching for potential studies for a review:

- searching by all methodology types yields studies that are more sensitive to the overall literature on the topic in question. However, this method of searching may identify studies that have less relevance (i.e. low specificity); or
- searching by specific methodology types yields fewer studies but these may be more relevant. (i.e. less sensitivity).

Searching by specific methodologies might therefore save time and resources but at the expense of introducing possible selection bias into the review.

Critical appraisal

Critical appraisal is an essential part of a systematic review. Explicit and transparent criteria are used to determine the quality and strength of the identified studies, and hence the weight attached to their findings. Studies which do not meet sufficient quality standards can be rejected. Example criteria that could be used to appraise studies using experimental designs are set out below:

- Question focus: was a clear and answerable question asked?

- Population/groups studied: were the populations and subgroups studied clearly reported, and was the sample size adequate?

- Selection bias: was there any selection bias in the achieved sample, and if so, was it effectively accounted for?

- Performance bias: were the trial and control groups treated similarly other than through the intervention?

- Statistical methods and reporting: were the statistical tests used appropriate to the questions beings asked, and were they reported adequately enough to permit validation and review?

Data extraction and organisation

A data collection form should be developed recording how, and why, data are to be extracted from named studies. A non-exhaustive and non-prescriptive list is set out below. The data that are relevant to the question being asked should determine the type of data extraction and organisation form which is appropriate.

- The nature of the interventions or processes studied;

- the studies' characteristics and methods used, the research design and analytical methods employed;

- the participants (populations and sub-groups) included and excluded; and

- the outcomes or processes measured/observed, and the main and subsidiary findings.

Analysis of data from sifted studies

The analysis of data from sifted studies will depend on the policy question(s) being asked, the type of methodology used in the primary studies, and the likely use to which the findings are to be put. Some issues to be considered in the analysis of included studies are suggested below:

- the appropriate comparisons (if any) to be made by the analysis, and the basis (study results) for making them – these might require some transformation or manipulation to make them comparable;

- the assessments of validity to be used in the analysis, and the analytical approaches to be used for making comparisons and summarising results, including meta-analysis;

- how heterogeneity/homogeneity of included studies will be dealt with; and

- the main findings of the review and the main caveats associated with the findings.

Summary and conclusions

Summary answers should be provided by a review as well as detailed analysis and conclusions to the policy question(s) being addressed. The review should be as clear as possible about what can and cannot be concluded from the existing evidence. It should also identify any weaknesses or limitations in the existing evidence on the topic in question. Finally, the conclusions of systematic reviews become less relevant over time as existing studies age and new studies become available, so any review should be dated and it's most recent update noted.

Rapid evidence assessment

6.20 Rapid Evidence Assessment (REA) is a pared down version of systematic review, employing the same general principles but in a lighter-touch manner to enable reviews to be undertaken more quickly. REAs collate descriptive outlines of the available evidence on a topic, critically appraise them, sift out studies of poor quality, and provide an overview of what the evidence says and what is missing from it. They are based on fairly comprehensive electronic searches of appropriate databases, and some searching of print materials, but not the exhaustive database searching, hand searching of journals and textbooks, or searches of the grey literature that go into systematic reviews.

6.21 Rapid Evidence Assessments carry a caveat that their conclusions may be subject to revision if more systematic and comprehensive review of the evidence is subsequently completed. This is consistent with the important principle that systematic reviews are only as good as their most recent updating and revision allows.

Meta-evaluation and meta-analysis

6.22 The term "meta-evaluation" was originally used to describe the "evaluation of evaluations" (Scriven, 1991) but has also been used to refer to "the synthesis of evaluations". It is similar to systematic review in that it tends to use explicit protocols and criteria for assessing the quality of evaluation studies. It tends to differ from systematic review in two ways:

- the evaluation will generally attempt to synthesise the results of the individual evaluations, either formally or informally, to provide some estimate of, for example, the average effect size across a range of similar studies, or the total combined effect of a number of related studies; and

- studies to be evaluated will not necessarily be identified through a systematic review of the entire relevant literature. Instead, they might be selected because they are of particular interest to the evaluation audience. This might be because they share a similar theme, were funded under the same programme, or were implemented in the same geographical area.

6.23 A meta-evaluation is relevant therefore where there are, for example:

- multiple policy interventions all working towards the same outcome, for example, interventions aimed at reducing childhood obesity;

- large scale programmes which have several strands with overlapping objectives, for example the legacy of the London 2012 Olympic Games covers economic, social and sporting impacts of the Games, as well as environmental and disability outcomes (http://www.culture.gov.uk/); and

- evaluations undertaken in different geographical areas using different approaches to achieve the same objective.

6.24 Meta-evaluation can use a range of more or less formal techniques for synthesising results and drawing conclusions. For instance, Meta-Evaluation of the Local Government Modernisation Agenda: Progress Report on Accountability in Local Government[10] used a range of techniques, including:

- a count of existing evidence reports with findings in favour of a particular result;

---

[10] *Meta-evaluation of the Local Government Modernisation Agenda: Progress Report on Accountability in Local Government,* Office of the Deputy Prime Minister, September 2006, http://www.communities.gov.uk/

- a questionnaire-based survey of local government officers; and

- focus group discussions with local residents.

6.25 As discussed elsewhere in the Magenta Book, the reliability of results obtained from techniques which use qualitative and other approaches which do not attempt to control for potential confounding factors is limited.

6.26 Meta-analysis is a more formal approach to meta-evaluation. It has been defined as "the statistical analysis of a large collection of analysis results from individual studies for the purpose of integrating the findings" (Glass, 1976)[11]. It is a type of systematic review that aggregates the findings of comparable studies and "combines the individual study treatment effects into a 'pooled' treatment effect for all studies combined" (Morton, 1999[12]). This can be based on a pooling of the individual observations from the original study datasets, but more commonly the average effect sizes estimated in each study are pooled. The variation in these effect sizes is then explained using statistical analysis, often multivariate regression using characteristics of the individual studies ("meta-data") as explanatory variables.[13]

6.27 Meta-analysis is perhaps best known for combining the results of randomised controlled trials, but they are also commonly undertaken on non-randomised data from primary studies that use case-control, cross-sectional, and cohort designs. Meta-analysis has its own limitations, including limits to the comparability of outcomes considered in different studies, and variability in the reporting of relevant meta-data. As with other meta-evaluations, the reliability of the results is a function of the quality of the "source" studies.

Making sense of existing and new evidence: simulation modelling

6.28 The outline logic model in Box 6.A at the beginning of the chapter is conceptually simple, but the examples presented in Box 6.B are in the most part quite involved, with each "step" in the logic itself implying a potentially large number of processes. For instance, the impact pathway for the health costs of air pollution involves complex physical, chemical, biological, technological and economic relationships between the generation of air emissions from electricity generation, meteorological conditions, human physical response to exposure to airborne pollutants, and individuals' attitudes towards changes in their respiratory health.

6.29 As suggested in Chapter 2, in cases such as these, it might not be realistic to expect even a well-designed evaluation to be able to detect any effect of one input – e.g. the amount of coal burned in a power station – and some ultimate outcome – e.g. individuals' health-related quality of life – in a single study. This is because there are too many confounding factors and too much "noise" in the pathway for the effect of one variable on a "distant" outcome to be detected. In these circumstances, an evaluation of a "shorter" set of links in the logic chain is likely to have more chance of producing a robust outcome (e.g. the effects of changes in air quality on reported respiratory health). However, there then remains the question of how the real relationship of interest (which might be the entire impact pathway) can be evaluated.

6.30 In other situations, reviews of the existing literature, using some of the techniques considered in this chapter, might reveal that there is a substantial body of robust evidence covering particular aspects of the logic model in question, but little or no evidence relating to others. This might mean that an evaluation which is restricted in scope and focuses on these less

---

[11] *Primary, Secondary and Meta-Analysis of Research*, Glass, 1976 Educational Researcher, Vol. 5., No. 10, Nov 1976

[12] *Systematic Reviews and Meta-Analysis, Workshop materials on Evidence-Based Health Care,* Morton, 1999, *University of California* San Diego, La Jolla, California — Extended Studies and Public Programs

[13] *The Essential Guide to Effect Sizes: An Introduction to Statistical Power, Meta-Analysis and the Interpretation of Research Result,* Ellis, 2010, United Kingdom: Cambridge University Press; *Practical meta-analysis,* Wilson and Lipsey, 2001, Thousand Oaks: Sage

developed areas will be considered more robust and better value for money than one that attempts to cover the entire impact pathway. The issue is then how the results of this new study can be combined with existing evidence to answer the evaluation questions.

6.31 Simulation modelling is one way in which the results of different evaluations of separate parts of the impact pathway or logic of an intervention can be combined. Simulation models are most commonly constructed in spreadsheet-style software using quantitative data. This requires that the evidence relating to the different links in the logic model are expressed in quantitative terms (e.g. effect sizes). It also means that the evidence must relate to comparable "endpoints", or at least to endpoints which can be "translated" into comparable measures. Box 6.G illustrates this using the example training intervention introduced in Chapter 2.

Box 6.G: Constructing a simulation model for a hypothetical policy intervention

Chapter 2 presented a (hypothetical) example policy to recruit unemployed individuals onto a new training scheme which provides seminars to improve work skills, with the intention of reducing the costs of unemployment.

A simulation model of a (full) economic evaluation of this intervention might require quantitative evidence on the following links of the implied logic model:

    1   measures of the resources used (costs) in delivering seminars;

    2   effect of the seminar series on (net) seminar attendance;

    3   effect of seminar attendance on participant skills;

    4   effect of change in participant skills on subsequent employment and earnings trajectories; and

    5   effect of changes in employment and earnings trajectories on quality of life and other relevant indicators (e.g. health status).

In this example, the endpoints of each stage in the logic model are the same, and hence are comparable, by construction. Evidence relating to each stage could therefore be linked in a simulation model with no need for "translation". However, if existing evidence relating to the fourth stage above was defined in terms of (e.g.) formal qualifications, but the evidence on the third stage measured skills in terms of specific abilities (e.g. reading and writing), then some translation might be necessary to estimate the "qualification equivalents" of the skill levels resulting from the intervention.

6.32 The example in Box 6.G suggests that some form of simulation modelling is likely to play a role in a large proportion of impact evaluations. For instance, where outcomes are expected to be affected materially over a number of years, some simulation of these effects might be necessary to ensure that evaluation evidence is obtained in a timely fashion. In addition, it might be difficult to detect in a single evaluation study an effect on lifetime earnings trajectories of attendance on a short-term training course at some point in the past, again suggesting the need to simulate any such effects (assuming there is evidence to support them). Finally, any wide-ranging economic evaluation will almost certainly require a simulation model, not least because many economic outcomes can only be measured through dedicated research exercises. An example might be a survey of affected individuals to estimate the value of changes in health status, which evidence suggests is associated with pollution-related reductions in air quality.

6.33 Whether a simulation-based approach to answering the evaluation questions will be appropriate and necessary is important to establish early on in the design of any new evaluation research study. This is because the need to use endpoints which are either comparable directly or can be translated into comparable terms might influence the design of the study, data

collection and so on. Selecting the incompatible outcome measures at the study design stage might make it impossible to make the necessary linkages in the simulation model, because there is no satisfactory "translation". Issues related to data collection are discussed further in Chapter 7.

6.34 Note that simulation-based evaluations will always be subject to some uncertainty about the validity of the assumed links and evidence underpinning them. With this approach, all outcomes are not measured directly, so the evaluation cannot "prove" that an impact was actually caused by the intervention in question. Where endpoints need to be translated to make them comparable, the translation will by necessity be based on assumption(s), and the validity of these assumptions will affect the reliability of the calculated impacts. In some cases, evidence relating to some links in the logic model might be relatively weak or even missing entirely, requiring stronger assumptions and introducing greater uncertainty. Many theory-based evaluations use significant amounts of qualitative evidence and assumptions to produce estimates of the impact of an intervention, and the uncertainty inherent in such information needs to be borne in mind when considering the reliability of the results.

7 Data collection

Introduction

7.1 Whatever evaluation approach is used, data will be required to evaluate a policy; what data will depend on the types of evaluation proposed and the research questions to be answered. There are four main types of data which, if planned for, might be able to play a key role in supporting both process and impact evaluations:

- existing administrative data that has not been collected specifically for the evaluation;

- long term, large scale, often longitudinal, structural survey data managed by central governments or the Office for National Statistics;[1]

- monitoring data or performance management data that are already being collected to support the administration of the policy; and

- new data collection needed to support the evaluations information needs.

---

[1] For example the Labour Force Survey, the British Crime Survey, the Wealth and Assets survey, the English Longitudinal Study of Ageing or the Birth Cohort Studies.

7.2 The availability of administrative or general long-term scale structural survey data should always be considered at the design stage of an evaluation because they have the potential to be important sources of background or explanatory data, for example unemployment rates used to explain crime trends.

7.3 In certain cases, where the evaluation has a sufficiently long lead in time, it might be possible to influence the collection of certain information through these sources, but this should not be relied on as a way to provide detailed project specific information.

7.4 As any administrative and long-term survey data will, by their nature, be being collected anyway this chapter will focus on monitoring data (which in some cases will be a sub-set of general administrative data relevant to the operation of the policy or programme), new data collection and data collection tools, before ending with a discussion about ethical and data protection considerations.

What is monitoring data and how can it contribute to evaluation?

7.5 Monitoring data can play a key part in policy evaluation by providing useful data to policy makers and analysts throughout the life of a policy. This can support both the monitoring of the policy as part of its routine management, and its evaluation (see how monitoring and evaluation fit into the ROAMEF policy cycle in Chapter 1).

7.6 Monitoring data are regularly collected about a policy and can include data relating to each component of the logic model (see Chapter 5 for further information on logic models) as summarised in Table 7.A.

Table 7.A: The types of monitoring data collected

Data	Example	Why collect this data?
The people accessing a service	Numbers and characteristics	This can help demonstrate whether a policy is reaching its target population
Inputs	Funding or staff numbers	This can inform a cost-benefit analysis and determine whether assumptions about the policy implementation, such as cost and time, were correct
Processes / activities	Referrals and waiting times	This can help determine whether the policy is being implemented correctly or whether there are any unintended consequences
Outputs	Numbers of people getting job interviews or number of applications processed	This can inform an assessment of whether the programme has delivered the target outputs to the anticipated quality
Outcomes	Employment rates and wages	This will help to measure the benefits of delivering the outputs

7.7 Monitoring data are frequently administrative and quantitative and are often not generated primarily for evaluation. However, this does not stop them from being a very useful resource for analysts, and the availability of this type of data, and whether there is any opportunity to adapt or collect it in a way that best support the evaluation should be considered at the design stage.

7.8 Monitoring data can provide answers to a number of policy, research and performance questions. Monitoring data may form the basis of an impact evaluation if the data is of sufficient quality and allows the estimation of a counterfactual. They also provide information to monitor the progress and performance of a policy from its start and can contribute to a process evaluation.

7.9 With reference to its role in supporting evaluation, monitoring data can be used to collect and measure data relating to:

- the logic model that underpinned the policy (see Chapter 5). Where, for example, an outcome (which may take some time to materialise) is dependent on a sequence of initial processes, if data show that these early stages are or are not happening this will have implications for the confidence policy makers will have in achieving their ultimate objective. For example, where a policy to reduce reoffending is thought to be dependent on an initial process of offenders regularly attending Probation services, and the monitoring data show a low rate of attendance, this data, in conjunction with the logic model may give an early indication that the policy is unlikely to be successful;

- the progress of a policy, programme or project against a set of pre-specified expenditure or output targets. For example the number of client contact sessions with the Probation Service against the target number of contact sessions;

- the numbers and characteristics of people, organisations and businesses accessing or using a policy. For example the demographics of offenders on a reducing reoffending programme and those who drop-out. This can help to determine whether the programme is reaching the target population and whether there are any differences among those that drop-out;

- the contact details of individuals, groups, organisations or agencies that are participating in or are subject to the policy and in some cases, the contact details of a control or comparison group. These can be used to inform the sampling strategies for follow-up research. Alternatively this data may be required to identify individuals on a further dataset, for example, to identify offenders on the Police National Computer to investigate whether they have reoffended;

- the impacts of a policy on central and local government and its agencies, such as hospital admissions and stays; arrests by the police and court prosecutions; enrolments in training course and university places; and use of social services and housing;

- the costs of a policy, this can include costs to other stakeholders, such as businesses or survey respondents, as well as government. For example, monitoring data may collect information on the amount of time Probation Officers spend on client contact sessions which can help calculate the total staff costs of implementing the programme or policy; and

- the economic effects of a policy, through changes in incomes, prices, employment, consumption and other economic measures and indicators of value.

7.10 Analysis of monitoring data can more generally help policy makers identify where a policy is not being implemented as expected and further action is required to ensure it can achieve its objectives. If the monitoring data suggests something is going wrong (such as fewer referrals to a scheme than expected), then policy makers or analysts may want to use an evaluation to check the extent of the "problem" and its reasons to inform contingency actions. Box 7.A provides an example of how monitoring data can be used within an evaluation.

7.11 From this it worth noting that care should be taken to establish the quality of the monitoring data being collected as poor quality or partial data will affect the scope and scale of monitoring data's contribution to an evaluation.

Box 7.A: An example of how monitoring data can inform an evaluation

Free Swimming Programme Evaluation (Department for Culture, Media and Sport)

The Free Swimming Programme began in April 2009 and was due to run to March 2011, but finished early in July 2010. It was funded by five government departments and was intended to get more adults, children and young people physically active. Funding was split into four pots: two supporting free swimming – one for 16 and unders, and another for 60 and overs, plus two capital modernisation pots. The evaluation had three main objectives:

- to measure changes in swimming participation;
- to identify lessons about what works, how, in what context and for whom; and
- to estimate the value for money, health and economic benefits of the programme.

A programme logic model was developed to provide a structure for the evaluation and guide the research. Evidence to measure the inputs, activities, outputs, outcomes and processes identified in the logic model was collected and analysed through a range of mechanisms:

- collection of monitoring data on the number of free swims and free swimming lessons from all 261 local authorities involved in the programme;
- analysis of the Active People Survey, a national c. 190,000 sample telephone survey which measures participation in sport and physical activity;
- a bespoke online survey of 4000 members of the population in the target age groups to assess participation in, and attitudes towards, swimming;
- case study visits to a sample of 12 participating local authorities;
- telephone interviews with a sample of 18 non-participating local authorities; and
- a literature review to assess the health and associated economic impacts of sport and physical activity.

A key focus of the analysis was on understanding the net impact of the programme. The key factors that impacted on the estimation of additionality[2] for this programme were:

- the reference case / counterfactual;
- deadweight (people who would have swum anyway, even if they had to pay);
- displacement / substitution (the extent to which the programme displaced swimmers from outside the target age groups, and how it impacted on participation in other sports);
- wider effects (the impact of the programme on paid swims by friends and family members); and
- sustainability (the likelihood of those who swam for free continuing to swim after the end of the programme).

The main evaluation findings were based heavily on the local authority monitoring data (for measuring gross impact) and the online survey data (for estimating additionality and net impact). There were concerns about data quality of some of the local authority data collection systems, but triangulation allowed an initial analysis for the revenue subsidy part

---

[2] The number of additional positive outcomes that the programme creates. It equals the number of positive outcomes achieved with the programme minus the counterfactual. It is a measure of the programme effect or impact. See Chaoter 6 for a more detailed discussion.

of the programme which suggested that the cost was greater than the benefit (in terms of avoided cost to the health service). The findings of the first annual evaluation report[3] informed the government decision to end the programme early in July 2010.

New data collection

7.12 All data collection, just like the policy allocation itself, needs to be planned before policy activity commences on the ground. This is to ensure that data are obtained about the baseline before the policy (or evaluation) started (this might be used in an impact evaluation as the counterfactual), as well as the situation once the policy is in operation. Consideration of what data are required, when they will be required and how they will be collected should be undertaken at the appraisal and implementation stage of a policy, Box 7.B covers the key areas to consider.

Box 7.B: The key considerations when planning for data collection[4]

What data need to be gathered to give reliable and consistent measurement against a policy's objectives?

What additional data should be collected to meet the policy maker's requirements for feedback on the policy and to support any planned evaluations?

Who will have responsibility for gathering data?

When will the data be gathered?

What are the key timeframes for collection?

How will the data be gathered, transferred and stored?

What format are the data required in?

How will the data be verified to ensure it is accurate and consistent with the relevant requirements?

7.13 Considering data requirements at the design stage of a policy has a number of benefits:

- policy makers and analysts can identify what regular information they need about the policy, the frequency with which they need it and ensure that this will be available to them throughout the life of the policy;

- data requirements can be designed into the policy so that it delivers this data as a routine process. This means that it can be costed and planned for by the people delivering the policy;

- baselines and counterfactual data can be collected; and

- where external organisations need to provide some of the data, the requirement to do this can be built into their contract (or Service Level Agreement or Memorandum of Understanding) from the outset – it may not be possible to add it later.

---

[3] *Evaluation of the Impact of Free Swimming*, PricewaterhouseCoopers, 2010 (http://www.culture.gov.uk/)
[4] *Evaluation Guidance Note,* Scottish Enterprise, 2008

7.14 The development of data collection plans should involve both policy makers and analysts, to ensure comprehensive coverage of all requirements and accuracy of research instruments and supporting policy descriptions. Where appropriate, it may also be useful for an external evaluation team, or the people who will deliver a policy, to contribute to the design process prior to the implementation of a policy. Where data for evaluation will be collected via monitoring data, the appropriate monitoring procedures and systems should be set up and embedded from the outset of an intervention, to ensure they systematically generate the appropriate data throughout the duration of the policy.

7.15 Final policy outcomes can take a long time to exhibit and so the collection of monitoring data must take into account the proposed time frames for each intervention. Where it takes too long to capture the final outcomes, or it is simply not possible to directly measure long-term outcomes it may be necessary to collect data on "intermediate" or "proxy" outcomes. Where this is the case these intermediate or proxy outcomes should be identified during the development of the logic model.

7.16 Careful planning for all data collection types is also required to ensure that ethical issues are fully considered, to account for the costs of data collection and to plan how data will be quality assured and transferred and stored.

Will monitoring data be able to be used in the evaluation?

7.17 Existing routine monitoring data has the potential to fulfil some or, on occasion, all the data needs for planned evaluation. If this is the case then the policy makers and analysts have the advantage of reduced costs, reduced intrusion upon operations and potentially a longer historical time frame in which to place observed changes in context. There are, however, clear limitations to these data in terms of what questions can be answered, and the data require substantial processing in order to be applicable to impact evaluations. Indeed this type of data may be of lower or higher quality than those collected expressly for research purposes, and are not independent; this should be considered when deciding whether or not to use it.

7.18 For example, information about exact dates of joining or leaving a programme are likely to be recorded accurately on monitoring systems, whereas the individual participant, if asked in interview, is unlikely to have perfect memory. But data about disability, for example, is likely to be more reliable when collected as part of a research exercise than through monitoring systems. For this reason it can be useful to collect data using a number of methods, such as monitoring data and bespoke surveys, this is known as triangulation and is covered in more detail in Chapter 8.

Who should collect the data?

7.19 In considering whether it is feasible for existing frontline staff to carry out the data collection task, analysts will want to consider issues such as:

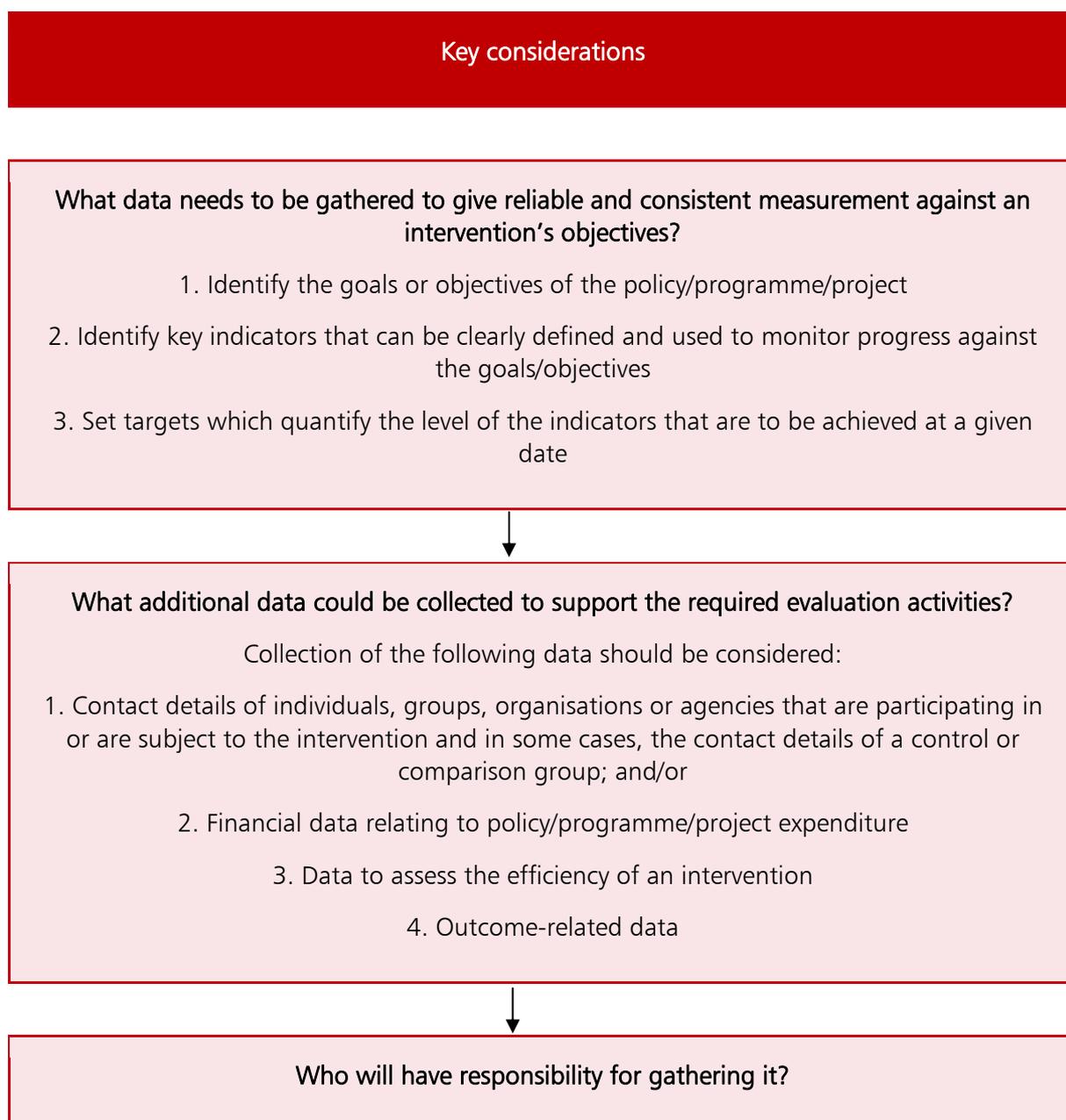- whether there is a culture that is open to research in the participating organisations;

- whether the participating organisations have a particular interest in a certain outcome;

- how heavily the new requirements would impact on the business as usual of frontline staff;

- whether frontline staff are well placed to know the information;

- whether using frontline staff will result in biased data;

- whether there is any means of verifying the completeness and accuracy of the data; and

- whether any necessary changes or additions to IT systems are feasible.

7.20 It is also important not to burden staff with a broad ranging request "for completeness" where there is not a clear match between the level of detail in the data being requested and the analyses actually planned. Indeed the researcher should be able to demonstrate how the data requested will enable the policy to be improved. Where in-house data collection is not feasible, or appropriate, potential alternatives include bespoke surveys, perhaps undertaken and quality-assured by internal or external analysts, and embedded research staff. It is worth noting that monitoring data can be distorted by changes in recording practices, for example, as awareness increases during the course of policy implementation, therefore it is important to ensure that data recording practices remain constant.

7.21 Box 7.C illustrates the key questions and considerations that need to be taken into account to design an effective monitoring system and subsequently to facilitate a good quality evaluation.

Box 7.C: Designing an effective monitoring system

Key considerations

What data needs to be gathered to give reliable and consistent measurement against an intervention's objectives?

1. Identify the goals or objectives of the policy/programme/project

2. Identify key indicators that can be clearly defined and used to monitor progress against the goals/objectives

3. Set targets which quantify the level of the indicators that are to be achieved at a given date

↓

What additional data could be collected to support the required evaluation activities?

Collection of the following data should be considered:

1. Contact details of individuals, groups, organisations or agencies that are participating in or are subject to the intervention and in some cases, the contact details of a control or comparison group; and/or

2. Financial data relating to policy/programme/project expenditure

3. Data to assess the efficiency of an intervention

4. Outcome-related data

↓

Who will have responsibility for gathering it?

| 1. Who is/are the most appropriate individuals to gather the data, e.g. programme/project delivery team, an existing performance monitoring team, and does this individual(s) have the capacity both in time and skills?

2. What resources are required to undertake the task? |

↓

When will it be gathered?

1. How often should the data be gathered, e.g. monthly, quarterly, annually etc?

2. Can the process be aligned with the auditing/reviewing process of the funding body?

3. Can the process be aligned with the reporting schedule for the evaluation?

↓

How will it be gathered and stored?

1. What format should the system use and can this be aligned with existing monitoring systems?

2. Data protection protocols to ensure the system is designed to meet security and data sharing requirements

3. What ethical e.g. informed consent and data protection considerations needs to be taken into account?

4. Where will the data be stored?

↓

How will the data be verified to ensure it is accurate and consistent with the relevant requirements?

1. Who is/are the most appropriate individuals to verify the data, e.g. analyst, programme/project lead at the funding body, independent evaluators etc?

2. What resources are required to undertake the task?

↓

Design and implement monitoring system

7.22 What if existing monitoring data is insufficient to answer the evaluation research questions?

7.23 Before launching new data collection processes it is important to review existing financial, administrative and monitoring data generation processes to identify whether the required evaluation data can be sourced from existing data sets, or an extension of an existing data set collection processes.

7.24 Frequently, however, new data, whether new monitoring data, or other forms of primary data, will need to be collected. This requires advance planning and ideally should be specified when designing a policy to ensure that the systems are in place to provide evaluators with the required data.

7.25 In the absence of regular data collection on the inputs, outputs and outcomes of a policy (which may be particularly important for an impact evaluation), subsequent evaluation may need to:

- highlight this as a shortcoming and identify the reasons for the data not being available; and

- take steps, as far as possible, to retrospectively collect and analyse data on the performance of the project.

7.26 However, attempting to retrospectively collect data in this way is not recommended. It is very likely to be more expensive than collecting data at the same time the policy was taking place. In addition, data may no longer be available or may be inaccurate or piecemeal and the opportunity to validate this data may have been lost. Information may not have been collected on drop-outs which may bias the findings. This is particularly relevant where this information is required to contact participants or where it is needed in order to identify them in other datasets. In summary, it can mean that an evaluation is not possible or that its findings are much less reliable than if data had been collected at the same time the policy was being delivered. Planning an evaluation, and its data requirements, early will therefore minimise the need to collect data retrospectively.

Designing data collection tools

7.27 Where monitoring data is not feasible or appropriate, bespoke research can be used to collect either process or impact evaluation data. This may be in the form of adding questions to existing surveys[5] (which are also useful for providing background information and as a source to sample from or weight back to), if timescales allow, or designing new primary research.

7.28 To meet the requirements of impact evaluations research will need to collect standardised data from both the treatment and control groups to allow for comparison against the counterfactual. Sampling must also be taken into account during the design of both quantitative and qualitative research to ensure that the sample size is large enough to achieve the desired information (for example, statistical power in a quantitative survey) to obtain robust results. These issues are considered in more detail in Chapters 8 and 9 and supplementary guidance.

Surveys

7.29 Surveys can be used to seek different types of information as covered in Table 7.B. However, it is worth noting that although surveys can be used to ask questions about behaviour it may not be the most reliable measure. Respondents may give socially acceptable answers (though good design and experienced interviewers can reduce this), or be genuinely uncertain about the true answer. For this reason, it might be necessary to observe behaviour rather than simply ask about it. Observation methods are discussed in more detail in Chapter 8.

---

[5] Either local or cross-government surveys such as the Labour Force Survey, British Crime Survey, British Social Attitudes Survey or the Family Resources Survey.

Table 7.B: The different types of information collected through surveys

Types of questions	The type of information collected
Factual questions	Surveys often offer the only practical and affordable way of collecting such information, and in some cases there is no other source or way of measuring the attribute of interest. This can include respondents' assessments of their health status, life satisfaction and so on.
Knowledge questions	Assess what respondents know about a particular topic and their awareness of the intervention being evaluated.
Attitudinal questions	Seek to measure respondents' opinions, beliefs, values and feelings which cannot be verified by reference to observation or external data sources.
Behavioural questions	Measure what people do or intend to do and how that has changed as a consequence of the intervention.
Preference questions	Respondents' preferences for different possible options and outcomes, including trade-offs between competing policy objectives. These can be used to elicit monetary values for different outcomes, including those not readily possessing market prices (e.g. changes in air quality, health status) for use in cost-benefit analyses.

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

- Can the respondents understand the question – and do they understand it in the same way that you do?

- Are respondents able to answer the question?

- Are they willing to answer the question?

- Will the question produce a reliable response?

7.31 Most data collection tools, whether qualitative or quantitative and their associated materials (e.g. show cards) will require developmental effort, possibly involving cognitive testing or pilots, to ensure they collect information effectively. This will be particularly true with complex questionnaires, such as those attempting to elicit preferences for social impacts, and further advice on writing and testing survey questions is provided in the supplementary guidance. It is important to ensure that new research is consistent with relevant existing or previous data monitoring and collection tools to enable comparison. Where possible, it is helpful to use standard formats for survey questions, or interview schedules, to ensure this consistency. This can have benefits not only for the particular evaluation study, but for building a wider evidence base, particularly where evaluations are being undertaken at a local level. There are some particular points to bear in mind when developing questions for quantitative surveys:

- the Office for National Statistics (ONS), in recent years, has been working towards harmonised questions for common variables such as age, gender, ethnic origin;[6]

- some questions have been extensively validated in previous studies – examples are the GHQ-12 set of questions for measuring mental well-being and the EG-5D questions for measuring health status.[7] Using these will enable comparisons with many other studies, and will ensure the results of the evaluation can be correctly interpreted;

- it may be appropriate in many cases to repeat surveys at the same time of year as a previous one, in the same geographical areas, or using the same sampling frame. A

---

[6] See http://www.ons.gov.uk/about-statistics/harmonisation/index.html
[7] See http://www.ons.gov.uk/about-statistics/harmonisation/index.html

good example of this would be evaluating crime reduction programmes. Some crime types are seasonal, for example, bicycle theft increases in the summer, in contrast burglary increases in winter. Therefore to be certain that a new programme was as effective as one previously evaluated, any data collection would need to be timed appropriately;

- in general there are likely to be tensions: between collecting precisely the ideal data for the current evaluation, and consistency with other studies; or between different, non-comparable, previous studies. To arrive at a balanced view, it is important to be clear as to which (if any) are the key studies to which the results are to be compared; and

- where possible, engage with specialist analysts within your own department or ONS when designing data collection tools.

7.32 It is also important to consider possible subsequent uses of the data during the planning stage and a particular consideration is whether the data can be archived, under what conditions and how much preparation it will need to ensure suitable anonymisation. A common practice is to make suitably anonymised data available through the Economic and Social Data Service.[8]

7.33 In some cases data may need to be kept for future use in such a way that individuals could be identified for future follow-up.[9] The informed consent process for any data collection will have to be designed with storage decisions in mind. Contracts with external research contractors will need to stipulate what outputs, including data sets, are to be provided and can include work to anonymise data. Contracts will also have to consider copyright of intellectual property including research tools as well as datasets.

Ethical and data protection considerations

7.34 Ethical and data protection considerations need to be taken into account when designing and undertaking any evaluation. However, the issues in these areas can be complex and sensitive, often requiring consideration on a case-by-case basis with analysts and other experts at the evaluation design stage and throughout the life of the evaluation. Best practice cannot cover all eventualities and so it is advisable to raise any areas of concern with the relevant Head of Profession/Senior Analyst.

7.35 It may also be necessary to gain ethical approval through an appropriate ethics committee, e.g. the Integrated Research Application System (IRAS),[10] the HSE Research Ethics Committee, the Social Care Research Ethics Committee etc. to undertake an evaluation. The need to gain this form of approval will depend on the content and form of evaluation being undertaken and should therefore be considered on a case-by-case basis. If an evaluation will involve research with vulnerable groups or individuals who lack the capacity to give informed consent, approval will need to be sought from an "approved body", for more information see the Department of Health factsheet for social scientists on the Mental Capacity Act.[11]

7.36 When considering data protection issues it will also be necessary to consider data security, transfer and sharing issues. This should include the consideration of non-disclosure and the physical aspects involved in data sharing (such as storing and accessing data) and in turn should

---

[8] http://www.esds.ac.uk/

[9] There are also Government Statistical Service protocols on data management, documentation and preservation – see http://www.ons.gov.uk/about-statistics/ns-standard/cop/protocols/index.html

[10] IRAS is a single system for applying for permissions and approvals for health and social care / community care research in the UK. See http://myresearchproject.org.uk

[11] *The Mental Capacity Act – factsheet for social scientists*, The Department of Health, September 2009

lead to the setting of clear data protection protocols which comply with the contractual arrangements of the relevant agencies.

7.37 There are a range of sources of information available to assist analysts and policy makers with ethical and data protection considerations and reference should be made to these, and other sources of guidance provided by the Civil Service professions, and any specific guidance issued by UK departments and devolved administrations when planning data collection.

7.38 Key sources include:

- The GSR code which, as an addendum to the Civil Service Code, sets out specific principles to guide the work and behaviour of Government Social Researchers, available at: http://www.civilservice.gov.uk/my-civil-service/networks/professional/gsr/

- GSR guidance on ethical assurance for Social Research in Government available at: http://www.civilservice.gov.uk/my-civil-service/networks/professional/gsr/

- The GSR ethics checklist which can help those designing or carrying out an evaluation to identify important issues to consider, available at: http://www.civilservice.gov.uk/Assets/gsr_ethics_checklist_tcm6-7326.pdf

- The codes of practice established by the Government Statistical Service – the UK Statistics Authority Code of Practice for Official Statistics (2009) http://www.statisticsauthority.gov.uk/

- The Social Research Association ethical guidelines http://www.the-sra.org.uk/

- The Market Research Society code of conduct http://mrs.org.uk/

- The British Psychological Society ethical guidelines and support http://www.bps.org.uk/

- The Data Protection Act 1998 http://www.legislation.gov.uk/

- The Freedom of Information Act 2000 http://www.legislation.gov.uk/

- The Freedom of Information Act (Scotland) 2002 http://www.legislation.gov.uk/

8 Process evaluation, action research and case studies

Key Points

- Process evaluation, action research and case studies can be used to evaluate the implementation and delivery of a policy to provide feedback on a wide range of issues. These can include whether the policy is being implemented as planned, what is working more or less well and whether it is delivering expected outputs and outcomes.

- Process evaluation cannot determine whether a policy "worked" this can only be achieved using an impact evaluation. It can, however, complement an impact evaluation by providing crucial insights into why a policy did, or did not, work and test the logic model on which the policy is based.

- It is important to consider at the planning stage the information requirements for any economic evaluation that process evaluation would be best placed to collect.

- Process evaluation, action research and case studies use a range of qualitative and quantitative research methods including one to one interviews, group interviews, surveys and observations. Multiple methods are often used to provide triangulation of data and corroborate findings.

Introduction

8.1 As has been discussed earlier, evaluation is not something that happens only after a policy has been implemented. Evaluation can be used throughout the life of a policy to provide policy makers with timely feedback about whether a policy is being implemented as expected, whether important outputs are being delivered and if there are any parts of the policy which are not working or which are working particularly well. Process evaluation, action research and case studies provide evaluation evidence on the implementation and delivery of policy which provides policy makers with the opportunity to refine and improve policies to help them have the best chance of achieving their ultimate aims. This chapter will describe the three evaluation approaches, presenting their similarities and differences, and describe the range of research methods used in these approaches, and key principles to consider when deciding which method/s to use.

Evaluation to understand the implementation and delivery of policy

8.2 A number of different evaluation designs can be used to understand the implementation and/ or delivery of a policy. As discussed in Chapter 5, choosing the most appropriate design will be dependent on a number of factors including the types of research question that need to be answered, how a policy has been delivered (e.g. national roll out or pilot) and practical issues such as when evidence is needed and what resources are available. The most common types of

research that might be used to evaluate the implementation and delivery of policy are process evaluations, action research and case studies.

Process evaluation

8.3 Process evaluation primarily aims to understand the process of how a policy has been implemented and delivered, and identify factors that have helped or hindered its effectiveness. It can take place at any time that the policy is being delivered (the timing of the evaluation will depend on the policy and research questions that need to be answered). Process evaluation can generate a detailed description of what interventions are involved in a service or policy, who provides them, what form they take, how they are delivered and how they are experienced by the participants and those who deliver them. It can also provide an in-depth understanding of the decisions, choices and judgments involved, how and why they are made and what shapes this. It can therefore provide timely information to answer the types of questions in Box 8.A.

Box 8.A: The types of questions answered by process evaluations

How was the policy delivered?

In what context was the policy delivered?

What did participants and staff feel worked in delivering the policy, why and how?

What did they feel worked less well in delivering the policy, and why?

What, therefore, might act as facilitators and barriers to desired impacts? How can barriers be overcome and facilitators harnessed?

Which particular aspects of the policy seem to have led to an observed outcome (in conjunction with an impact evaluation)?

Was the policy implemented "on the ground" in the way it had been planned? (This could include observation of the "take up" of a service or policy, or "compliance" where the policy includes regulation or legislation. It could also include identification of unintended outcomes.)

How consistently was the policy implemented across multiple sites or did local variations mean that effectiveness was diluted?

Did the policy meet its targets for inputs and outputs? (To establish the need to investigate causes of any difference between expectation and delivery.)

Was the logic model (see Chapter 5) linking policy and outcomes supported in the experience of the people delivering or receiving the policy?

Did recipients and staff understand the intervention?

What was the experience of recipients and staff who received and delivered the intervention? Which aspects were most valued or caused difficulties? Was this different for different groups of people?

What was the nature of interactions between staff and recipients during the roll out?

Who did not engage, or dropped out, and why?

How effective were risk management strategies in anticipating and mitigating risks?

Did the policy meet budgetary expectations when rolled out, or were there unforeseen issues and hidden costs?

How might the policy be refined or improved?

8.4 Process evaluation can therefore provide information to assess how a policy is performing, improve the quality of the policy, if needed, and inform future policy development. As noted, this information can be important in explaining the results of an impact evaluation. In particular, without a process evaluation it may not be possible to assess whether a policy that appears not to have had an impact is actually flawed itself or has been affected by poor implementation and delivery. Additionally, where a policy is shown to "work", a process evaluation might indicate which elements of the policy appear to be most influential and therefore how resources might be most efficiently used. It might also supply the data which can be used in an impact evaluation to test the influence of different aspects of the intervention. Where local contexts appear to have influenced the success of a policy this can also help policy makers consider how likely results are to be duplicated in other situations and circumstances.

8.5 There is no single way to conduct a process evaluation. The evaluation can be designed to meet the exact information needs of a particular policy. The methods, timeframes and costs will all therefore depend on what information is required. It is very important to be clear about the aims and objectives of a process evaluation by:

- identifying all the policy questions the research will need to answer and when these will need to be answered;

- drafting clear research questions that reflect these policy priorities;

- identifying what data will be needed to answer these questions: including who will have this information, which groups will be studied (and what sampling techniques will be needed if not all participants will be included), what format it will be collected in (for example. paper or electronic), and when the data will be available;

- deciding at what stage in the policy a process evaluation will provide most value (if not throughout the study); and

- understanding how the resulting evaluation will support an assessment of the policy's performance, refinement of the policy, or an impact evaluation.

8.6 Process evaluations might collect and analyse quantitative or qualitative data to answer the research questions, or a combination of both. It is important that whichever data is used, that it is collected accurately, analysed robustly and presented appropriately. This could mean identifying an appropriate and credible sampling technique to choose research participants, ensuring appropriate statistical techniques are used when quantitative data is analysed, or choosing a range of methods or groups of participants to corroborate findings or deepen understanding (also known as triangulation of data). Triangulation of data, or the use of multiple methods, which explore similar research questions adds credibility to and confidence in the findings of an evaluation and strengthens the conclusions and recommendations that can be made as a result (triangulation is discussed further in Table 8.A below).

8.7 Input and outcome measures can feed into any economic evaluation, so it is important to consider the information requirements for cost-benefit analysis when planning a process evaluation. Often, the process evaluation might be the principal or even the only source of additional data for an economic evaluation. Therefore, if the special data requirements for economic analysis are not considered when designing the process evaluation, a meaningful economic evaluation might be effectively precluded, as it will not be possible to collect the information retrospectively.

8.8 Further information about methods that can be used to support process evaluations (and action research and case studies) and how they can be chosen to best answer the research questions set for an evaluation is discussed below.

Table 8.A: Types of Triangulation (Denzin[1] 1989)

Methodological triangulation	This refers to combining different research methods. This can include "within research" triangulation (where, for example, a range of different lines of questioning might be used to approach the same issue) and "between method" triangulation (where different data collection methods are combined). This can also include the combining of qualitative and quantitative data.
Data triangulation	This means combining data from more than one source, for example, a number of settings, points in time or groups of people.
Investigator or analyst triangulation	This involves more than one researcher looking at the data so that they can either check or challenge each other's interpretation or deliberatively approach the data from different angles.
Theory triangulation	This means looking at the data from different theoretical positions in order to explore the fit of different theories to the data and to understand how looking at the data from different assumptions affects how it is interpreted.

Action research

8.9 Action research is an approach to evaluation which can help policy makers and practitioners make changes to improve policy at an early stage in policy development and increase the likelihood that a policy will achieve its aims. Action research involves the researcher and those involved in developing and implementing the policy collaborating to diagnose actual problems and develop solutions based on this diagnosis.[2] To maximise the benefits of action research, this collaboration should be very active and this type of research is likely to require a lot of input from both researchers and policy makers. Action research often coincides with a policy's implementation to identify issues that might occur at this stage ensuring that implementation is as effective as possible, and anticipating and addressing any issues that arise at this early stage. However, it can successfully be used at all stages of the policy process.

8.10 Examples of when action research might be particularly useful are where:

- a novel way of working or delivering an intervention is being implemented;

- a policy is based on a new or unproven theory of change (see Chapter 5 for more information on theory of change) and little evidence is available about how it might work in practice;

- there are a number of feasible alternative options for delivering a policy and it would be helpful to test them; or

- a policy is being delivered in a challenging implementation environment.

8.11 However, action research is well placed to meet a wide range of policy needs, particularly when a quick, responsive, problem solving approach is required.

8.12 Action research is likely to require collection of data to understand the environment in which a policy is being implemented or delivered and data to diagnose any problems with this process. It also needs to collect data to help identify possible solutions to improve the policy or its delivery. This might include collection of quantitative and qualitative data or a combination of both (possible methods of data collection are discussed below). The key aspect is that the action

---

[1] *The Research Act: A Theoretical Introduction to Sociological Methods,* Denzin, 1989, Englewood Cliffs, NJ, Prentice Hall.
[2] See *Social Research Methods,* Bryman, 2001, Oxford: Oxford University Press

researchers should regularly feed back their analysis of this data to the policy maker and/ or practitioner and together they should identify key problems and possible solutions. If possible, it is helpful for the action researcher to further evaluate these changes to the policy to ensure that they are having the desired effects.

8.13 Therefore where action research is being carried out, it is important that policy makers and/or practitioners are willing to make changes to the policy as a result of the action research as its value is in changing how the policy is being delivered on the ground. It is particularly important to consider when it would be most appropriate for data collection for any impact evaluation to take place. It is desirable to begin this data collection once the action research has been completed so that only the impact of the improved policy is captured, otherwise a possible finding that the policy has little, if any, impact would be of no value for future decision making.

Case studies

8.14 In this chapter case study is defined as an in-depth, possibly longer term investigation of a single or very limited number of people, event, context, organisation or policy. A case study might be used when seeking to understand a significant or novel situation and to provide particularly rich data. Although the conduct of a case study can sometimes appear to be similar to that of a process evaluation, including in the generation of research questions and choice of methods (discussed below), there are key differences between the two which will affect how they are conducted and how the data generated can be understood.

8.15 Case studies will tend to be more localised or context specific than a process evaluation. That is to say they may look at a small-scale policy or project that is happening in only one, or a very small number of areas, and with limited numbers of participants. The policy or event being studied may even be a one-off situation such as the impact of the 1980 Cuban expulsion of workers into Miami on the labour market.[3] Alternatively, a case study approach may be used to investigate a larger scale policy but the case study itself would tend to focus on the experience of the policy for a limited number of people or in a limited number of locations which are of particular interest to policy makers. This type of case study may be used on its own simply to provide data about the people or areas of interest or may contribute to a larger process evaluation by providing this more in-depth account as part of a wider analysis of the overall implementation and delivery of a policy. Whatever the context, case studies are likely to be used when what is required is a very detailed, in-depth understanding that is holistic, comprehensive and contextualised.

8.16 Case studies will tend to use a variety of research methods and triangulation (see Table 8.B) to develop a clear, well reasoned and comprehensive understanding of the situation, project or people being studied. This can provide very useful learning for analysts and policy makers to identify why something happened or did not happen, the mechanics of how a policy works, how people worked together or how behaviour was influenced, and very in-depth information about how a policy has been working in practice. As a result, this can help generate hypotheses and templates for wider roll out of a promising policy or suggest ways of working that might work in similar circumstances. However, because of their focus on a limited number of examples, unique situations, or small-scale projects, case study data should not simply be generalised to a context beyond that being studied and it is important that their results are reported and used with this understanding.

8.17 In some situations a case study approach can be used to assess both the implementation and the results of policy in a particular area, generating quantitative data to support an

---

[3] *The Impact of the Mariel Boatlift on the Miami Labor Market*, Carol, 1990, Industrial and Labor Relations Review Vol. 43, No. 2 (Jan., 1990), pp. 245-257

evaluation of the policy's impact. In this situation, if the data is suitably robust, there is sufficient sample size and an appropriate comparison group to assess what would have happened in the absence of the policy (the "counterfactual"), then an impact evaluation may be possible within the case study. Guidance on impact evaluation can be found in Chapter 9. However, there is likely to be a lower possibility of generalising findings from impact evaluations which have been conducted within a small case study than in evaluations which have a broader scope.

Why undertake a process evaluation, action research or case study?

8.18 These three types of research have many overlaps, being able to answer similar research questions and tending to use a similar range of methods to collect data. However, they do have different principal characteristics which are presented in Table 8.B.

Table 8.B: The principal characteristics of process evaluation, case studies and action research

Type of research	Characteristics
Process evaluation	Probably the widest and most flexible of the three types of evaluation. Investigates a number of different research questions to understand and chart the implementation and delivery of a policy. Summation of past activity (whilst still having the aim to influence and improve future practice).
Action research	Interactive and iterative research which is used to influence the development of the policy being implemented. Therefore involves close collaboration between the researchers and policy makers. Requires commitment from policy makers to swiftly and continuously reflect upon and amend their policies, which may not be feasible with large scale policy implementation.
Case studies	Focussed on smaller scale or more localised aspects of policy delivery providing a level of detail that would be unwieldy if replicated for the full breadth of standard policy implementation. Might be used as part of a wider process evaluation.

8.19 These types of research can also be used in combination to strengthen the evaluation of a policy's implementation and delivery. For example, action research might be undertaken when a policy is initially being implemented to refine its procedures and practice, and a process evaluation could then assess the delivery of the final version of the policy. Alternatively, a case study approach could be used within a process evaluation to provide more detail and in-depth data on a context, area or situation of particular interest.

8.20 Taken together, these types of research may be particularly useful when:

- evaluating a new or innovative pilot project where rich data is needed to evaluate what has worked more or less well - including how a policy might be streamlined and made more efficient and how it might be developed in order to be rolled out to a wider audience;

- assessing best practice to identify aspects of policy delivery that appear particularly effective or successful in the area(s) being studied and which might provide a model for similar work in similar areas;

- identifying how to develop or improve service and policy delivery (for example, the evaluation of the Sure Start children's centres showed that there were barriers to

fathers participating fully, and was able to give useful suggestions as to how this could be improved);[4]

- investigating local variation and practice and whether this has a positive or negative influence on implementation;

- assessing/ identifying unintended or unforeseen consequences of the policy that might affect the overall impact of a policy; and

- conducting an impact evaluation will not be possible or will be severely constrained. This might include a small-scale project where the sample size is too small to support an impact evaluation, a project that is rolled out nationally so there is no opportunity to create a comparison group, or a policy where the impact of interest may not be measureable or cannot be measured until too late in the policy cycle. Monitoring data or process evaluation in these situations could provide descriptive data of performance against agreed targets or outputs and qualitative assessments of efficacy.

8.21 These types of research can also supplement and complement an impact evaluation with rich data to explain the impact (or lack of impact) that has been observed. Evaluation of the implementation and delivery of a policy can specifically help explain why, how and for what reasons policy outcomes occur, whereas impact evaluations tend to focus on what, where and when outcomes occur.

8.22 For example, a process evaluation may identify that a policy has not been targeted correctly, (such as a community service intended for the socially deprived that has actually been primarily accessed by more affluent and established members of the community) which means that the expected outcomes were unlikely to occur. Alternatively, it could explain why the intended recipients of a policy have not engaged with it or why the policy has not met their needs (for example, a service to get people into employment may initially have successful outcomes but if the employment is not suitable for their skills or existing commitments people may resign).

8.23 Importantly, a process evaluation can provide further data to explain differences observed in an impact evaluation. For example an impact evaluation might show more or less impact for different groups of service recipients and a process evaluation or case study could provide insight into their experiences of the policy which might explain these variations in success.

8.24 Process evaluations, action research and case studies can therefore answer a range of policy and research questions and are very flexible and useful analytical tools. As discussed in Chapter 5, as with all evaluations, however, in order to get most benefit from them, it is important for policy makers and analysts to identify what specific information will be needed about a policy at the design stage. This will help identify what type of evaluation will be most appropriate and effective and at what stage(s) data should be collected and analysed. Box 8.B provides an example of a process evaluation.

---

[4] *Fathers in Sure Start local programmes*; Lloyd, O'Brien and Lewis, 2003, NESS Research Report 04,DfES; HMSO. http://www.education.gov.uk/publications/

Evaluation of provision of calorie information by catering outlets (Food Standards Agency)

Provision of nutrition information in catering settings, specifically calorie labelling, formed a part of the previous government's wider programme of activities to tackle a range of diet related public health issues, including obesity. In January 2008, the Food Standards Agency (FSA) launched an initiative beginning with the voluntary provision of calorie information (CI), at point of choice (POC), as the first step to providing consumers with more consistent nutrition information when eating outside of the home.

The aim of the evaluation was to explore the practical implications for the 21 businesses participating in the pilot scheme in setting up and running the scheme and to get an early understanding of consumers' (respondents who took part in group discussions and who indicated that they regularly ate in the types of catering outlets represented by the participating companies) and customers' (respondents who took part in observations and interviews in the participating catering outlets) use and understanding of the scheme, to provide information on what worked and where improvements could be made. A process evaluation approach was adopted and several different methods were used.

Business research

- 39 business interviews (20 Head Office, 19 Outlet Manager) were conducted in person or over the phone depending on businesses' preferences – exploring why the business participated in the research, how they set up the scheme, decisions around display of the CI and how issues during set up and roll out were dealt with.

Customer research and consumer research

- 289 customer interviews across the country in catering outlets; 143 POC observational interviews where behaviours were observed and consumers asked about how they were choosing their food; and 146 post choice interviews shortly after people had made their food choices – explored understanding and use of CI in purchasing decisions and views on presentation of CI.

- Eight group discussions with consumers in four locations – explored in more detail issues which were raised in the customer interviews.

The main findings of the evaluation were:

- participating businesses were generally positive about their involvement in the pilot and most set-up issues were overcome with relative ease; there were some concerns about further roll out that would need to be addressed (e.g. ensuring adequate IT systems in place);

- actual usage of CI was low, but consumers could envisage ways in which CI might be used (e.g. balancing meals); and

- the capacity and inclination of consumers to use the information, was dependent on three factors: visibility (presentation of CI should ensure that the text stands out so that it is noticed), understanding (additional information, e.g. reference information, is helpful for consumers to interpret CI accurately) and consumer engagement (the use of positive messages when displaying information helped

engage consumers with CI).

The findings from the evaluation were used to develop proposals for a voluntary calorie labelling scheme and these were put out to consultation in early 2010 and have shaped and set guiding principles for the scheme.[5]

Research methods to support process evaluation, action research and case studies

8.25 There is no single way to undertake process evaluations, action research or case studies. They are very broad types of evaluation design in which analysts, in consultation with policy makers, can choose a variety of methods to answer the particular research questions, considering the timescales, resources and data available. As well as considering the immediate questions that a policy maker might want to answer about the implementation and delivery of a policy, it will also be important to consider whether an impact and/or economic evaluation will also be conducted. If so, then in the planning stage, consideration should also be given to what data might be required to inform, and explain the results of, the impact evaluation and how the delivery and implementation evaluation could collect relevant data.

Choosing research methods

8.26 When designing a process evaluation, case study or action research the principles in Table 8.C should be followed in translating research questions into the range of particular research methods that might be used.

Table 8.C: Principles to consider when selecting research methods

Principles	Explanation
There must be a clear set of research questions that can be addressed through the delivery and implementation evaluation.	Research questions that are broad or vague can easily lead to unsatisfactory studies that simply do not produce new insights or do not have sufficient relevance or reliability to aid future decision making, which means that the evaluation will not offer value for money.
There should be coherence between the research questions and the populations and data studied.	Populations and data that are going to give the most direct and insightful information on the subject matter should be selected, taking into consideration which subsets of these populations are critical for inclusion or exclusion. (More information on sampling is provided in supplementary guidance). For example, researchers and policy makers may want an overall assessment of how a policy has been experienced/ implemented for everyone receiving it, but also to understand if there were different issues/ experiences for different genders, ages or ethnic groups. If there is already robust evidence on the experience of a particular group of service recipients, then an evaluation may want to focus on gathering data on other recipient groups rather than duplicating previous research.
Building comparisons into the design can be helpful and lead to more in-depth understanding.	For example, a study looking at a particular phenomenon among lone parents (such as attitudes to work) might be enhanced by including couple parents. Comparing the responses of the two groups will help with understanding of what is a function of being a lone parent, as opposed to that of simply being a parent.

---

[5] *An evaluation of provision of calorie information by catering outlets, prepared for the Food Standards Agency,* BMRB Social Research, December 2009, http://www.food.gov.uk/

There should be coherence between the research questions and the settings studied.	For example, sites should be chosen to provide coverage of the populations of interest to the policy makers. This could range from specific locations, organisations, contexts or groups of people all the way to collecting national data.
There should be a logic between the research questions and the data collection methods used.	For instance, are naturally occurring data needed because what is being researched is best illuminated by observing behaviour or interaction? (This might be the case where there is reason to believe that people's self-reported behaviour might not reflect what actually happens in practice.) Or do the research questions require evidence of people's own experiences, opinions and views? In which case data might be best collected through individual interviews or group discussions. Alternatively, if quantitative data (for example statistics on service take-up) is required then this might be most appropriately met by using existing monitoring data or commissioned surveys.
There should be a logic to the timing of the episodes of data collection.	This would include deciding at what stage of delivery and implementation information should be collected, and if data is required at a number of intervals. For example data may be collected to assess levels of attendance on an employment course at the start, middle and end of the course, or assessments of educational achievement may be made of a group of students before and after they receive a new educational intervention.
It is important to consider the feasibility and appropriateness of a proposed methodology within the actual research setting.	For example, it would be important to check that researchers would be allowed to observe particular aspects of service delivery, such as counselling sessions, before adding this technique into the evaluation design.

8.27 It will also be important to consider at an early stage the criteria against which a policy or service is to be evaluated, and what data will credibly demonstrate if these criteria have or have not been met. Table 8.D provides a list of questions to use as a guide when designing process evaluations, case studies and action research. The questions should be asked for each research question to ensure that they drive the study design and choice of methods. The same method can be used to answer a number of questions and this should be taken into account when designing the research tools and sampling.

Table 8.D: Key considerations when designing process evaluations, case studies and action research

Key question	Considerations
What type of data will be required to answer each research question?	• Is numerical data required? • Is factual (documentary) data required? • Is observational data required? • Is data to describe people's experiences, opinions, and views required? • Will a combination of these types of data be required?
Who or what can provide this data?	• Which participants, service providers, stakeholders, databases etc., would have this data and/ or need to be consulted? • Do/ will researchers be able to get access to this data? • Are there any potential sensitivities/ ethical issues in collecting data from these groups, areas, databases etc.?
What section of the population of interest should data be collected from?	• Who is the population of interest? • Will the research be a census of all available data/ population of interest or will a sample of the population be studied? • For qualitative sampling – what range of people, experiences, organisations, contexts etc. need to be covered? • For quantitative data, what types of estimate will the data need to provide and how precise? Which sub-populations need to be included? What impact does this have on the sample size required? • For qualitative and quantitative data – what sampling frames are available or will need to be created?
How will the data be collected?	• Which research method is best placed to provide the required type of data from the required sources (see below for summaries of key research methods)? • Is the data already being collected or will new data collection be required for the research? • When should/ can the data be collected? • How will data be validated/ triangulated? • Who will collect the data?
How will the data be analysed?	• Does the method of analysis that will be used require a particular sample size or type of data to have been collected?

Research methods

8.28 The Magenta Book does not provide detailed guidance on how to design and conduct individual research studies using different methods. However, some of the methods most commonly used in process evaluations, action research and case studies are briefly introduced in Table 8.E (further information on the qualitative data methods discussed is provided in the supplementary guidance).

Table 8.E: Research methods used in process evaluations, action research and case studies

Interviews	Interview data can provide rich information about the attitudes, opinions and experiences of people involved in a policy to provide in-depth information about how it is working in practice. They allow participants to explicitly explain their views, decisions or actions, describing what has shaped them. Interviews with key participants can be structured (a set list of questions is used with all interviewees), semi-structured (a list of questions with flexibility to probe further) or unstructured (no set list of questions). Interviews most commonly take place face to face between an interviewer and one interviewee, but might also take place over the telephone. The key people to interview will vary from policy to policy but may include those implementing a policy (including a range of levels of seniority and job roles), those receiving a policy, and also stakeholders with an interest in the policy. Usually the analysis of interviews is based on examination of the content, but less frequently techniques of conversational analysis can be used to analyse the *way* that things are said, by looking at speech patterns and/or body language.
Group interviews	Group interviews provide an opportunity to collect information for a group of people on their attitudes, opinions, perceptions and experiences, building and reflecting on each other's ideas and suggesting a variety of viewpoints and proposals. In group interviews data can be shaped through group interaction. Group interviews can be used with the range of people delivering or receiving a policy. They can work very well in tackling abstract or conceptual topics, where on a one-to-one basis a participant might "dry up". In group interviews, the researcher usually acts as a facilitator and works to a core script which sets out key questions or issues to be discussed by the group. Group interviews can work well in combination with one-to-one interviews or other research techniques. For example, at the beginning of a study they can be used to understand people's current practise, behaviour and beliefs, and test understandings of issues that can then be investigated later in one-to-one interviews. At the end of a study, they offer a deliberative forum for examining the implications of the study's findings for service delivery or policy development, and/or generating or prioritising solutions, with a focus on providing practical suggestions to improve the policy or service. Group interviews can be particularly useful with research participants who may find one to one interviews "scary".[6]
Observation/ participation	Observing or participating in a policy as it is being delivered provides researchers with direct experience of how a policy is working in practice, for example, a researcher may observe court hearings or benefit interviews. Data will tend to be recorded by the researcher either in narrative form or in a pro-forma, at the same time as the intervention they are observing/ experiencing or shortly afterwards. (In practice, most observational research is non-participatory.)

---

[6] *Focus Groups in Feminist Research,* Madriz, 2000, in Denzin and Lincoln (eds) Handbook of Qualitative Research. Sage: Thousand Oaks

Surveys	Survey or questionnaire data provides structured, often quantitative data on people's attitudes, opinions and experiences. It may be possible to repeat surveys to map changes in these factors during the life of the policy. This means that surveys can provide statistical data to understand the people, organisations and areas affected by a policy at one or a number of points of time. Depending on how the survey is set up, this can provide data that can be generalised to the whole population of interest. Surveys may be administered in a number of ways including face to face, telephone, internet and postal, each of which has positive and negative implications with regard to issues such as response rates and cost. Questionnaires are most often used to collect quantitative data but can also contain free text questions to collect qualitative data. When designing a survey it is important to consider what sample of participants and what type of analysis will be needed to answer the research questions and this should be built into the evaluation design. (Further information on survey design is provided in supplementary guidance.)
Consultative and deliberative methods	This describes methods that are used for consultative purposes (for example by local government). Boundaries between consultative research and other types of qualitative research are not absolutely clear cut, and some consultative methods involve the application of established research methods to situations where issues are being debated or deliberated. These types of methods will tend to be used when analysts and policy makers want to go beyond exploring people's views and behaviours, to getting them to come up with, or appraise, solutions and strategies. A wide range of public participation methods might be used including meetings, interactive websites, citizens' panels and juries, deliberative polls and participatory appraisal. Consultative research generally involves intensive exercises with relatively small groups, and thus raises questions about value for money and representativeness. However, well-conducted consultative research will help to highlight and explain areas of difference, as well as agreement, among participants. A careful balance therefore needs to be struck between the need for consultative research to identify an agreed way forward and the danger that it produces an artificial consensus.
Statistical analysis of quantitative data	A number of sources of quantitative data (including administrative and monitoring data, survey data, and numerical case file data) can provide statistical data on a policy's delivery that is very useful to a process evaluation. For example, quantitative data may be used to calculate numbers of participants receiving an intervention, their characteristics and initial information about costs.
Document analysis	Access to and analysis of documents relevant to the policy being evaluated can provide rich data on all aspects of the policy, including direct commentary on it by those involved in its implementation. These might include computer records, case files, referral letters, diaries, pictures etc. These data can be collected and analysed using appropriate content analysis techniques.

Ethnography	Ethnography is a method used by anthropologists which has been adopted by social researchers more generally. It is the detailed description of a culture, group or society, and of social rules, mores and patterns around which that culture, group or society are based. Ethnography is able to elicit the cultural knowledge of a group or society and also involves detailed investigation of patterns of interaction within it, in order to understand the values, processes and structures of that group. Ethnography tries to study social groups and activity in as 'natural' a way as possible. Observation, listening, remembering and detailed note taking are key techniques for researchers using ethnographic or participant-observation methods of inquiry. Amongst other benefits, this type of data can provide robust evidence on how front-line agencies work, identify variations in the social and cultural environment within which policies, projects and programmes are expected to work, and key personnel who might operate as "product champions" for policies, programmes and projects.

8.29 Whichever research methods are used it is important that the collection, analysis and presentation of data for process evaluation, action research and case studies follows best practice. This should include consideration of sampling strategies where appropriate and an understanding of how the achieved sample (the range and characteristics of the people or organisations that took part in the research and the amount of non-response and missing data there was) will affect the presentation of findings (for example how tentative or firm conclusions should be and how the sample is described). Guidance on sampling for the collection of qualitative data is provided in supplementary guidance. It should also inform how far, if at all, findings from a study of a particular policy can be generalised. Particular issues for sampling in qualitative research are discussed in Box 8.C.

8.30 Analysts should also reflect on the quality of data that has been collected (particularly when utilising monitoring data that has not been generated specifically for the evaluation), and also whether chosen methods of analysis are appropriate to the data collected and to answer the research questions. Whilst these issues are noted here to aid reflection on how to present findings from implementation and delivery evaluations they are issues that should be borne in mind for all types of evaluation and research.

Box 8.C: Key principles for sampling in qualitative research

Qualitative research sampling has a quite different logic from that of quantitative research. The objective is to select the individual cases that will provide the most illuminating and useful data to address the research questions. The intention is not to provide a precise statistical representation of the research population but to reflect aspects of its diversity which are expected to generate insight. The two main approaches are:

- **Purposive sampling**: sample cases are chosen deliberatively to represent characteristics known or suspected to be of key relevance to the research questions. These selection criteria are set at the first stage of evaluation design, based on existing research, expertise, or hypotheses. The composition and size of the sample is then determined and individual cases selected to fit the required composition.

- **Theoretical sampling:** in this case the researcher makes decisions about the type of data to collect and participants to involve next as the study proceeds, on the basis of emergent theory from their analysis of initial data.

Qualitative samples need to be large enough to include key subgroups and to reflect diversity. The emphasis is on mapping and understanding issues, rather than counting or numerical representativeness. In fact, large samples can be a hindrance as data gathered in qualitative research are rich and intensive. Depth lies in the quality of data collection and analysis, not quantity. The appropriate size of a sample will vary and is always a matter for judgment, but it also needs to be reviewed during fieldwork and as fieldwork draws to a close so that gaps in sample coverage can be filled. The same principles apply for group data collection methods, such as group interviews. Finally, the sample frames used in qualitative research are varied, as in quantitative research, and might include existing data sources such as survey samples, administrative records, registers or databases, or sources which are generated specifically for the research.

8.31 In summary, process evaluations, action research and case studies can use a range of methods, both quantitative and qualitative, which provide important information about how a policy has been implemented and delivered. They cannot, however, conclude whether a policy was successful or not, this can only be captured through impact evaluations, as discussed in Chapter 9.

9 Empirical impact evaluation

<div style="border: 1px solid red;">

Key points

- Empirical impact evaluation seeks to find out whether a policy caused a particular outcome to occur. It requires both a measure of the outcome and a means of estimating what would have happened without the policy, usually using a comparison group.

- Empirical impact evaluation is not feasible for every policy, especially if there is no comparison group. It may also be constrained if data are not available, or are too noisy, on the things it is necessary to measure.

- Impact evaluations cannot be guaranteed to produce the correct answer. There is always some risk of concluding that a programme works when it does not, or that it is ineffective when it has a real impact. To some extent the risks can be mitigated by careful design of the research, and sufficient investment in data collection, but they also depend on, among other factors, the actual size of the impact.

- The comparison group may have different outcomes from the policy group because of the way it was selected, rather than because of the policy itself, making comparison "unfair". This problem is known as selection bias.

- Research designs seek to control the composition of the comparison group so that selection bias can either be avoided or taken into account. Using randomness plays a central role here, but this does not always mean a randomised control trial. Sometimes "natural" randomness present in the system being studied can be utilised instead.

- The analysis of evaluation data requires an "identification strategy" to isolate the policy effect from competing influences. This involves modelling the sources of selection bias either directly (for example, by regression) or indirectly (for example, by estimating their effects with respect to trends over time).

- Reporting of an evaluation should distinguish between descriptive statistics on the outcomes and true impact evaluation, which takes potential non-policy causes for observed changes into account. The former cannot answer the question of whether the policy caused the observed changes to occur, but the latter can.

</div>

Introduction

9.1 This chapter focuses on impact evaluations which provide a quantitative measure of the extent to which any observed changes in an outcome of interest were caused by the policy. This kind of evaluation attempts to estimate the counterfactual – that is, what would have happened to the outcome of interest had the policy not taken place – by controlling for other factors which might have caused the observed outcome to occur. The outcomes can be selected to answer a range of questions, from whether the policy achieved its ultimate objectives, to whether other, intermediate outcomes were affected, which might indicate how and why such

changes occurred. (The latter questions are also discussed in the context of process evaluation in Chapter 8).

9.2 The scope of this chapter is confined to empirical methods which isolate the effect of the policy from other factors affecting the outcome of interest through the statistical analysis of newly-collected or existing data. It does not, therefore, consider those types of impact evaluation which attribute changes in an outcome to the policy (or aspect of it) through reference to theory or existing evidence (this is discussed in Chapter 6).[1]

9.3 The formulation and analysis of the research designs used in impact evaluation require a solid grounding in statistics, and often expertise in a range of specialised techniques. The supplementary guidance provides a more detailed explanation and technical treatment. This chapter is therefore more concerned with the concepts, rather than the mechanics, of impact evaluation. To present these concepts it makes reference in places to particular research designs and statistical techniques, and as such is slightly more technical than the rest of the Magenta Book. But this is not a "how-to" guide to those techniques; rather, it seeks to explain carefully the underlying issues that arise in impact evaluation and what the techniques can and cannot do to address them. It should be useful both to analysts seeking to advise their policy colleagues on setting up evaluations, as well as to those responsible for managing externally-commissioned research as critical customers.

9.4 This chapter begins by considering what is required to conduct an impact evaluation, why it is sometimes problematic, and under what circumstances it is feasible. The next section builds on Chapter 3 and looks at the fundamental principles behind designing policies for evaluation, and how they may be applied. The important issue of "noise" is then considered. A section on data analysis follows, built around the notion of an identification strategy. The different ways in which research designs attempt to address selection bias are discussed, and some of the things that can go wrong are considered, along with advice on detecting and correcting for them where possible. Finally, there is a section on "constrained designs", including guidance on reporting results when the evidence falls short of what would be regarded as acceptable for a full impact evaluation.

Introducing empirical impact evaluation
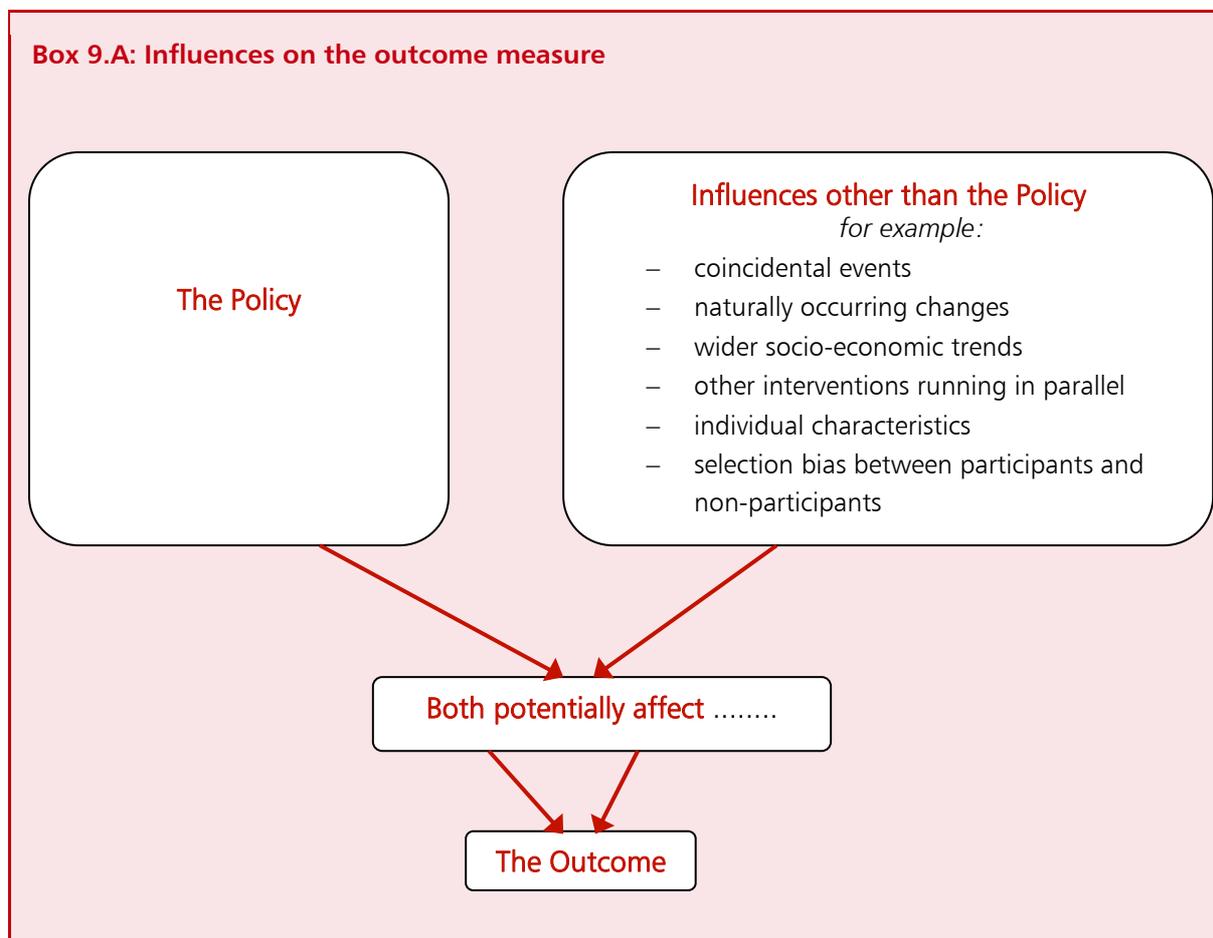
9.5 Fundamentally, evaluating policy impact involves:

- determining whether something has happened (outcome); and

- determining whether the policy was responsible (attribution).

9.6 The first of these points lies in the realm of descriptive statistics and is an important first step which has its own challenges. But it is the second point – establishing attribution – that is the defining feature of impact evaluation. This second stage is frequently the more challenging of the two, and can restrict the types of policies for which impact evaluation is feasible. The main problem is that other causes outside of the policy might have affected the outcome, as illustrated in the influence diagram in Box 9.A. The challenge of impact evaluation is to separate the effects of the policy from the other influences.

---

[1] The rest of this chapter uses impact evaluation to mean empirical impact as defined in 9.2

Box 9.A: Influences on the outcome measure

The Policy

Influences other than the Policy
for example:
- coincidental events
- naturally occurring changes
- wider socio-economic trends
- other interventions running in parallel
- individual characteristics
- selection bias between participants and non-participants

Both potentially affect ....

The Outcome

9.7 A key concept in impact evaluation is the **counterfactual** – what would have occurred had the policy not taken place. By definition it cannot be observed directly, because the policy did take place. Impact evaluation seeks to obtain a good estimate of the counterfactual, usually by reference to situations which were not exposed to the policy.

9.8 In broad terms, a robust impact evaluation requires:

- a means of estimating the counterfactual;

- data of adequate quality and quantity to support the estimation procedure; and

- that the level of "noise" in the outcome is sufficiently low to detect what might be a reasonably expected policy effect.

9.9 In practice, some or all of these requirements may be outside the control of the evaluator. To meet them often requires putting measures in place before the policy starts. For example, this could include manipulating the allocation of interventions (discussed below and in Chapter 3), and setting up appropriate data collection both to act as a baseline and during the policy intervention.

9.10 The remainder of this chapter is largely concerned with **research designs**, typically involving a **comparison group** as a means of estimating the counterfactual. But in some very simple cases, the mechanism may be sufficiently transparent that the impact can be observed directly, or through process evaluation, without the need to control for confounding factors. For example, with a project to supply water to a village in a developing country, any observed decreases in the average time household members spend collecting water might be attributed to the project

without the need for a comparison group[2]. As suggested in Chapter 2, the more "distant" are the factors or links in the logic model between which it is desired to estimate the impact, the more likely it is that there will be a range of possible explanations for any change in the outcome of interest, and the more important it will be to estimate a counterfactual. More often in public policy, the causal link between policy and outcome is an indirect one, and a counterfactual estimate is required.

When are empirical approaches possible?

9.11 The requirements mentioned in the previous sub-section cannot be met for every policy, so quantitative impact evaluation is not always an option. It may therefore be necessary to manage expectations around policies for which impact evaluation is less feasible, particularly if the policy is small scale and the additional data collection required to evaluate it would be too difficult or expensive to undertake. Box 9.B summarises the features of policies that are likely to make empirical impact evaluation either more or less feasible. These features are discussed in more detail in the remainder of this chapter; their relative importance depends on the individual policy, so not every feature is necessary for every evaluation. It is important to note that cases cannot be separated simply into "possible" and "impossible", as set out below, there are finer gradations in between with some cases being more or less likely to yield valid results.

[2] *Some Reflections on Current Debates in Impact Evaluation,* International Initiative for Impact Evaluation (3ie): Working Paper 1, White, 2009, New Delhi

Box 9.B: Circumstances affecting whether empirical impact evaluation is feasible

	MORE FEASIBLE IF…	LESS FEASIBLE IF…
Scale of impact	Direct relationship between outcome of interest and driver whose effect it is desired to assess	Complex ("distant") relationship between outcome of interest and driver of interest, with many potential confounding factors
	Large effect relative to other changes taking place is expected	Small effect is expected
	Effect is realised within a short time period (and does not vanish immediately thereafter)	Effect builds up gradually over an extended time period
Data availability: what was done where when to whom outcomes	Policy involves a distinctive change in practice with respect to identifiable subjects (individuals, institutions or areas)	Policy involves a consolidation of existing best practice, or is poorly differentiated between subjects
	Data available on individual subjects	Only coarsely aggregated totals available
	Data available on precise time periods	Uncertainty over timing of implementation (requires aggregation over time)
	Data to support evaluation collected before and during policy	Data to support evaluation not sought until policy already established
Potential comparison groups	Pilot undertaken at the start including data collection in non-policy areas	No pilot, or data available only for the pilot areas themselves
	Phased start across areas	Simultaneous launch nationwide
	Objective allocation, for example using a cut-off score or random allocation	Subjective allocation
	Accidental factors influencing allocation	Optimal targeting: a "perfect" allocation can frustrate impact evaluation by leaving no equivalent comparison group

Designing policies for effective evaluation

9.12 This subsection begins by introducing the theory behind research designs. A key part of successful impact evaluation is ensuring that a group of individuals or areas unaffected by the policy – the untreated – can serve as a comparison group. Such a group can be constructed in numerous ways, and several examples will be considered; these examples could form a basis for discussions between policy makers and analysts at the policy design stage. A separate sub-section, below, develops some of the concepts further as they apply to the analysis of the data obtained.

9.13 It is worth noting that the methods of allocating policies described in this sub-section all rely on there being something tangible to allocate. That is, the policy needs to consist of specified interventions such that it is possible to say distinctly that some individuals or areas were exposed to them, and others not (and further, that there is no impact on those who were not exposed). The methods of this chapter are not well suited to evaluating higher level-strategies, which set out aims and principles for action, unless those strategies can be unpacked into their constituent activities. The first task for the evaluator when faced with that kind of evaluation problem is to ascertain how the strategy is to be implemented: what will interventions look like on the ground, and who will receive them.

Randomness

9.14 Randomness[3] plays a central role in establishing the counterfactual to a policy. Randomness in the way policies are administered can balance out unobserved (sometimes, unobservable) differences in characteristics between the treated and untreated groups. The groups are then said to be equivalent – they differ on average only in their exposure or not to the policy. Comparisons between equivalent groups are said to have strong **internal validity** [4]: the evaluator can (under particular circumstances) infer that any significant differences between the two groups were caused by the policy, because on average the two groups are similar in all other respects.

9.15 The difficulty with evaluating actual policies is that they tend to target the most problematic or deserving individuals, institutions, locations and so on. That is, policies tend to be non-random intentionally. So even when one group is exposed to the policy and another is not, the two groups will typically be non-equivalent. Drug treatment policies, for instance, target individuals with drug misuse problems, who are likely to be different from other people in quite particular ways (for example they are more likely to be younger, male, unemployed and with an offending history than people who are not drug misusers). Allocation of the policy or intervention is then said to be endogenous to the outcome which is being targeted, because the characteristics which make an individual (or area or business) more likely to receive the intervention are also likely to affect impact of the intervention on their outcomes. Estimates of the policy effect which do not take this into account will suffer from selection bias, and simple comparisons between the treated and untreated groups are not then valid.

Research Designs[5]

The purpose of research designs is to manipulate the implementation of the policy, or to exploit features which it already possesses, in such a way that a counterfactual can be estimated. Manipulating the policy is preferable because randomness can be introduced, or non-

---

[3] Randomness" is used here in its widest sense, of events occurring by chance. "Randomisation", where a chance mechanism is introduced into policy allocation quite deliberately, is an important special case, but is not the only way in which randomness can occur.

[4] Internal validity and external validity are two terms often used to describe the strength or otherwise of an evaluation design. They can be explained by reference to the evaluation of a programme piloted in a small number of areas. Internal validity is where we can estimate the impact on the people who took part in those areas; external validity is where you would get the same impact in other areas, or at another time

[5] This chapter of the Magenta Book uses the term "research designs" to include both experimental and quasi-experimental designs.

randomness addressed, by design. Otherwise, a successful evaluation might need to rely on the required characteristics appearing by accident, and this is by no means guaranteed to be the case. So how should a good comparison group be obtained in practice? There are two approaches which will be considered in turn:

- Experiments, or Randomised Controlled Trials (RCTs). The defining feature of this approach is that the assignment of eligible individuals (or areas) to treatment is explicitly randomised, as it were by the flip of a coin.

- Quasi-experimental designs (QEDs). These designs do not use explicit randomisation, but address potential non-equivalence of the treated and untreated groups in other ways.

Randomised Controlled Trials (RCTs)

9.16 An RCT is usually regarded as the strongest possible means of evaluating a policy, because of its ability to balance out the differences between the groups. As was pointed out above, policy allocation by its very nature is not usually random, so opportunities to use it in practice are limited. If the policy is by intention "experimental", however, then randomised allocation might be more readily acceptable. In these instances the policy will usually begin with a pilot in a restricted number of areas only.

9.17 Randomisation can face some practical hurdles in a social research context mainly rooted in the difficulty in maintaining complete control over the allocation process, and the near impossibility of "blinding"[6] for the sorts of interventions being tested in public policy. It may get excluded because of (sometimes unfounded) concerns over ethical issues[7], or because an "experimental" design is at odds with a desire to focus the efforts of the policy in a targeted way. Both these arguments presuppose that the intervention is effective in the first place, which it is the purpose of the evaluation to ascertain (unless strong existing evidence already supports it – in which case the value of a pilot, randomised or otherwise, might be arguable anyway). In the latter case it may still be possible to incorporate randomisation for a limited subgroup of eligible participants. Boxes 9.C and 9.D provide two examples of randomised control trials.

---

[6] Blinding" refers to feature of experiments in which neither participants, nor those interacting with them, are aware who is in the treatment group and who in the control group. This is most easily understood in the context of drug trials, where it is necessary to guard against the well-know placebo effect, whereby somebody who believes they are getting an improved treatment can respond positively regardless of whether there is any direct effect. To overcome this, treatment and control group members receive apparently identical treatments, and have no way of knowing which they are receiving. Further, because those monitoring their progress may – consciously or unconsciously – record results differently for those they know to be receiving the alternative treatment, they also need to be 'blind' to the allocation. In social policy experiments, this is extremely difficult to achieve. For example, if the 'treatment' was a course of training, it would be readily apparent to all who was receiving it and who was not.

[7] Sometimes, perhaps because it is less common as a means of evaluating social policies, it is supposed that choosing who will benefit from a pilot intervention by random allocation is somehow unfair or unethical. Yet it is no more unfair than allocating treatment on the basis of where somebody lives, which is a much more familiar process.

Box 9.C: An example of a randomised control trial

Evaluation of HM Prison Service Enhanced Thinking Skills programme (Ministry of Justice)

There is considerable international evidence, from various systematic reviews and meta-analyses analysing a large number of offending behaviour/cognitive behavioural programmes, to support the effectiveness of these programmes in reducing re-offending. However, the evidence from research in England and Wales on the effectiveness of these programmes is mixed. This project looked at a shorter-term impact than reconviction to assess the efficacy of the Enhanced Thinking Skills (ETS) programme in the UK.

The main aim of the project was to examine the impact of ETS courses on 'impulsivity' in adult male offenders over the age of 18, and to investigate whether changes in levels of impulsivity were reflected in changes in prison behaviour. Impulsivity, a behaviour targeted for change by ETS courses, was chosen as the main outcome measure as there is research evidence of links between impulsivity and offending (e.g. Mak, 1991, Eysenck and McGurk, 1980).

Further analysis of individual cases was undertaken to investigate evidence of reliable clinical change. A secondary aim was to explore a range of other psychometric measures in the ETS test battery to evaluate the wider effectiveness of ETS courses, and to examine background factors of offenders, and institutional factors, in order to determine which offenders benefit most from ETS programmes, under which conditions. This was to see whether there were improvements to be made in course content, targeting of offenders, and selection of the most appropriate assessment methods.

A Randomised Controlled Trial (RCT) was selected in order to minimise bias in allocation of participants to groups. However, RCTs have rarely been conducted in UK prisons, largely due to ethical concerns about withholding treatment from a control group. These concerns were avoided by adopting a waiting list control design in which all eligible offenders ultimately received treatment. Offenders with a priority need to attend a course were assigned to a parallel cohort group prior to the random allocation, and their data were analysed separately.

However, it is not possible to assess the impact of the ETS course on reoffending through this study as all participants eventually received the intervention (hence there was no control group for reoffending analysis).

The study demonstrated positive results with regard to the (short-term) effectiveness of the ETS programme. More specifically, the study revealed that ETS programmes are effective in reducing both self-reported impulsivity and the incidence of prison security reports in adult male offenders.

Additionally, the analysis of background factors raised a number of issues relating to which offenders benefit from ETS programmes and how others may be assisted to benefit more. This could lead to better targeting of offenders for ETS courses, and adaptation or development of programmes specifically designed to meet different needs. The evaluation also raised questions about the relationship between offence type, impulsivity and effectiveness of ETS courses with different offence groups, which may lead to a greater understanding of particular types of offending and ways to reduce offending.

For more information read the evaluation reports online.[8]

---

[8] *Evaluation of HM Prison Service Enhanced Thinking Skills Programme,* McDougall, Perry, Clarbour, Bowles and Worthy, 2009), Ministry of Justice Research Series 3/09 http://www.justice.gov.uk/publications/

Box 9.D: An example of a randomised controlled trial

Primary School Free Breakfast Initiative (Welsh Assembly Government)

The Welsh Assembly Government made a commitment to introduce free healthy breakfasts in primary schools in Wales from September 2004. By January 2007 all primary schools had been offered the opportunity to participate with more than 1000 schools involved. The coalition Government's 'One Wales' commitment of 2007 was to maintain the programme.

 A cluster randomised controlled trial, with an embedded process evaluation, was commissioned in May 2004 to assess the impact of providing free breakfasts in schools on children's eating habits, concentration and behaviour. The cluster randomised design was chosen because randomisation at the individual level was not possible as the programme was implemented at the whole school, rather than individual pupil, level. The cluster randomised approach is often chosen for settings based interventions, such as schools or workplaces.

The study recruited 111 primary schools, of which 56 were randomly assigned to the control condition and 55 to the intervention. Data were collected at each for three time points: baseline, four month and twelve month follow-up. In each school, one Year 5 (age nine to ten years) and one Year 6 (age ten to eleven years) class were randomly selected, resulting in a repeated cross-sectional survey of approximately 4350 students at each data point.

The evaluation team concluded that the results provided partial support for the scheme as a dietary intervention. The 12 month follow-up found that:

- 41 per cent of pupils in intervention schools that had started a scheme attended at least once a week, with 30 per cent of these attending each school day;
- the quality of breakfasts eaten improved among pupils in intervention schools, with consumption of items such as fruit, vegetable and wholemeal bread increasing;
- more positive attitudes towards breakfast were found in intervention schools;
- there was no significant effect on breakfast skipping, episodic memory or inattention; and
- the absence of a decrease in breakfast skipping was suggested to be unsurprising, given the relatively small number of breakfast skippers at baseline. The evaluation team recommended that further work be undertaken in promoting pupil uptake and reach to address the breakfast skipping issue.

There is an existing evidence base suggesting that breakfast consumption influences cognitive functioning and classroom behaviour. The lack of impact on cognitive functioning in this study is likely to reflect the fact that this was analysed at school level, influenced by uptake, rather than tracking change at the individual level.

For more information read the evaluation reports online[9]

---

[9] *An Evaluation of the Welsh Assembly Governments Primary School Free Breakfast Initiative*, Murphy, Moore, Tapper, Lynch, Raisanen, Clark, Desousa, and Moore, November 2007, http://www.wales.gov.uk

Quasi Experimental Designs (QED)

9.18 Suppose, however, that randomisation has for whatever reason been rejected. A QED should then be considered. Fundamentally, these designs use one of two approaches (or sometimes, a combination of both):

- exploiting natural randomness in the system to obtain a comparison group that is "as good as random", insofar as group membership does not depend on any factors likely to affect the outcome; or

- acknowledging that the comparison group is non-equivalent, but obtaining it in a way that allows selection bias to be modelled (typically in some form of regression model).

9.19 Some of the options for obtaining a comparison group are shown in Table 9.A. It is worth mentioning that phased introduction is arguably the most robust approach of those listed, and if full randomisation is deemed unsuitable then this approach should always be given serious consideration at the policy design stage.

Pilots

9.20 Designing evaluation for a pilot involves essentially the same considerations as for a larger scale policy, but there are some additional caveats:

- If the pilot is on a very small scale, its effects may not scale-up as expected. There could be greater enthusiasm among those involved with the initial piloting than would be encountered more widely. The dynamics of administering the intervention could be rather different among a small group than would be the case with more widespread implementation. Therefore, unless the pilot is simply a proof-of-concept it should try to operate through the same administrative structures as will be used in an eventual wider policy.

- Piloting can provide the evaluator with a ready-made comparison group in the form of areas similar to those where the pilot took place, but not operating it. However, unless the evaluation uses only administrative data, it will be necessary to carry out data collection in the comparison areas as well. That could be more problematic as staff working in those areas will face an additional burden from taking part in the evaluation, without gaining the potential benefits of early assignment to the new policy. An alternative is to allocate treatment and control groups within a pilot area.

Addressing non-randomness

9.21 Whether the comparison groups in Table 9.A are "as good as random" depends on the details of how they arise, or are constructed, for any particular policy. For example, if a phased introduction is used and the assignment of areas to waves is essentially arbitrary (or indeed, has actually been randomised) then it is reasonable to compare areas that are in the first wave with those that are not. On the other hand, if the highest priority areas are placed in the first wave, then the comparison group must be regarded as non-equivalent, and selection bias is a real possibility. Another issue is that consistency of delivery may change over time, especially if the first wave embraces the new policy more enthusiastically than the later waves.

Table 9.A: Example sources of a comparison group

Phased introduction	The policy is phased-in in "waves" rather than introduced simultaneously in all geographical areas. During the period when not all areas are implementing the policy, the areas assigned to the later waves can form a comparison group for the earlier ones. This is similar to piloting but can be more rapid, as there is no presumption of an evaluation being completed on the first wave before the second is launched. It does however require that the impact occurs on a short timescale, relative to the interval between waves, and that the details of the policy do not change between waves. It also assumes that behavioural effects and impacts are not triggered with the policy announcement.
Intermittent application	If the policy involves interventions that are very short term in nature (such as media campaigns, for example) then applying these in intermittent bursts, where different areas receive them at different times, can be used to compare active areas to quiet areas. Once again, the impact needs to occur on a short timescale if this approach is to be used.
Accidental delays	Policies that begin simultaneously nationwide are problematic with regard to area-based studies. But it is worth investigating whether for practical reasons some areas went ahead more rapidly than others. If a frank account of the degree of implementation can be obtained from each area, a comparison group of "slow starters" might emerge. If there is a "postcode lottery", the evaluation can make use of it.
Intensity levels	If simultaneous introduction of the policy is unavoidable, another strategy is to evaluate based on differing modalities or intensities in different areas. Where there is local discretion on how the policy is implemented, it may be possible to classify different areas according to the decisions they made; where some areas receive enhanced funding or run additional interventions, these areas may be compared with those operating only the basic policy. In both cases, however, the impact estimated is for the difference between variants of the policy rather than for the policy as a whole.
Administrative rules	A comparison group may arise as a result of having to "draw a line" to decide who receives an intervention. For example, an offender aged 17 years 11 months may be very similar to one aged 18, but treated completely differently by the criminal justice system.
Targeting	Whenever a policy is intended only for a certain subpopulation (of individuals or areas), those unaffected by it form a potential comparison group. Almost always in this scenario, the comparison group will be non-equivalent.
Non-volunteers	Where participation in a programme is voluntary, those who do not participate can be a source of a potential comparison group. Such a comparison group will always be non-equivalent and controlling for the differences will be challenging.

9.22 So, if the comparison group is not "as good as random", what can be done about it? At the policy design stage, the points to consider are:

- how allocation to treatment will occur (whether intentionally or accidentally) and how this might lead to non-equivalence;

- what data can be captured on the known characteristics of individual subjects, for use in subsequent analysis; and

- whether it is possible to design the policy so that allocation uses an objective rule, based on these known characteristics of those who might be targeted. If it can, then evaluation will be stronger, because the sources of selection bias are all known about.

9.23 The topic of modelling selection bias is developed further in the sub-section on data analysis below.

9.24 In relation to the third bullet above, a special case of an objective allocation rule is to form an "assignment score" based on the level of need of each individual. Those above a certain score receive the intervention. An elegant method of analysis is then offered by the regression discontinuity design (RDD; supplementary guidance will provide more detail on RDD). This design is based on examining the boundary between the "only just eligible" and the "not quite eligible". The scores (both of participants and non-participants) need to be captured for future analysis. The main drawback of the RDD is that the results only apply directly to those at the boundary, and may not be an accurate indicator of the effects on individuals with characteristics away from the threshold.

9.25 Voluntary participation in an intervention is an example of non-randomness that is a particular problem for the evaluator. It is tempting to use individuals who opted not to participate in some scheme (or chose not to complete the course) as a comparison group for those who did, but the fundamental flaw with this approach is that opters-in are very likely to be different from opters-out, and in particular are likely to be better motivated. Motivation might be important if it is a significant determinant of the effectiveness of the intervention (for example educational courses being more effective with motivated students). This "self-selection" is another example of a non-equivalent comparison group, and can be one of the hardest to address. Some possible solutions are:

- attempt to control for motivation directly. However, motivation is difficult to observe by nature and standard administrative data such as demographics about the prospective participants are unlikely to capture it. Therefore, specialised surveys may be required in an attempt to elicit participants' reasons for the decision, and this can be a costly exercise. Alternatively it may be possible to find proxies for motivation. For example, studies on schemes to help non-employed people into work[10] have found that previous labour market history gives a good indication of motivation, if recorded in sufficient detail;

- carry out the analysis on the basis of intention to treat (ITT). The policy group consists of all those offered the intervention, even those who decline, and a comparison group is drawn from individuals who would have been eligible but were not offered (perhaps because they were associated with an institution that did not operate the scheme at the time)[11]. Impacts estimated on an ITT basis tend to

[10] *The econometric evaluation of the New Deal for Lone Parents,* Department for Work and Pensions Research Report No. 356. 2006. http://www.dwp.gov.uk/

[11] *The econometric evaluation of the New Deal for Lone Parents,* Department for Work and Pensions Research Report No. 356, 2006. http://www.dwp.gov.uk/

be smaller than those based on an actual treatment group, since the ITT group is diluted by non-participants, and it may not be possible to distinguish the impacts from the "noise" (see below). However this approach can have stronger internal validity and is arguably more policy relevant, since it measures the effect per person of making the policy available, which can actually be controlled; and

- examine what happens downstream of the decision to participate. If some individuals who consented were later unable to participate due to unavailability of resource or other administrative reasons (but not due to reneging, which would reintroduce selection bias) then these individuals can provide a comparison group.

Power of design[12]

9.26 Selection bias arises from underlying differences between the treatment and comparison groups, which might cause them to have different outcomes irrespective of the policy. Bias affects all members of a group, on average, in the same direction. For example, with an urban redevelopment initiative the treatment areas might be more deprived than the comparison areas. The success with which a research design is able to address these systematic differences is called the strength of the design. Strength is a subjective concept and is not a numerical quantity.

9.27 In addition, there are also random differences between individual members of both groups which affect their outcomes independently. For example, some pupils taking a school test might do well just through luck or less well due to "having a bad day", irrespective of underlying ability. These kinds of differences appear as random fluctuations or "noise" in the outcome measure. The power of a design is its ability to detect policy effects in the midst of "noise". Power is a numerical quantity – it is defined as the probability that if the true effect is of a given size, then the design will detect it with a given level of confidence, or at a given "significance level".[13] The relationship between power and strength is shown in Table 9.B.

Table 9.B: Experimental power vs strength

	Weak design Poor counterfactual or none at all	Strong design Realistic counterfactual estimate
Low power Small number of observations and / or policy effect small relative to noise	Unlikely to detect difference between groups or over time. And even if we do, we have no confidence in attributing it to the policy.	Unlikely to detect difference between groups. But if we do, then we have confidence in attributing it to the policy.
High power Large number of observations and / or policy effect large relative to noise	Very likely to find a significant difference between groups but this does **not** mean it can be attributed to the policy.	Very likely to find a significant difference if there is a real policy effect. We have confidence in attributing this difference to the policy.

9.28 Power depends both on the size of the effect on the outcome relative to the natural variation in that outcome (or the "signal-to-noise ratio") and on the number of observations. It also depends on the research design being used. As an illustrative example, Box 9.E is concerned

---

[12] This section assumes a basic knowledge of statistics, for example hypothesis testing and the t-test.

[13] Significance is a function of the "noise", or variance in the outcome of interest. If the change in an outcome is said to be "significant at a five per cent level", it means that, given the natural variance in that outcome, a change of such a magnitude would only be expected five per cent of the time.

with the power of a simple test of difference between two means (based on an unpaired t-test)[14] as might be used to analyse the results of an RCT. It shows the number of observations required to achieve a power of 80 per cent at a significance level of five per cent for a range of signal-to-noise ratios. What is quite striking is that if the size of the policy effect is similar to or greater than the noise, then quite small sample sizes (e.g. 15 treated and 15 controls to give a combined sample of 30) are adequate; but as the relative signal size decreases, the number of observations required to detect it increases dramatically. For example, a signal-to-noise ratio of 1:8 would require a combined sample size of 2000.

Box 9.E: Sample size requirements for a simple t-test

Signal: Noise	Total N
4:1	6
3:1	8
2:1	12
1:1	34
1:2	130
1:4	500
1:8	2000
1:25	20,000
1:100	300,000

The table shows the combined sample size (treatment + comparison group) required for an unpaired t-test if it is to have a power of 80 per cent at a significance level of five per cent. The "signal" is the mean treatment effect and the "noise" is the residual standard deviation.

9.29 Is it possible to predict the signal-to-noise ratio, and hence the required sample size, in advance? The expected noise level may be estimated from historical data if available, but the signal – that is, the predicted policy effect – is trickier. It may be possible to estimate it from the logic model of the intervention, reasoning along the lines of how many people will be affected and what might be a realistic change in their behaviour as a result. It may alternatively be possible to calculate how big an effect would need to be in order for the policy to be considered a success (either in political or cost-benefit terms), and to say that if the actual impact was less than this it would not matter if it was undetected.

9.30 The implication is that impact evaluation is only worth attempting on policies where the expected impact is large enough to stand out from random fluctuations in the system under study. How large is large enough depends on how well modelling is able to explain the differences between individual group members that arise in the absence of the policy. If it is possible to predict accurately what an individual's outcome "should" be, then any impact on that outcome due to the policy is easier to detect. If, however, the drivers of these differences are poorly understood, or are not captured in any model, then the noise level will be higher. Small schemes or minor refinements to practice that may still be good value for money and entirely worthwhile on the basis that "every little helps" cannot then have their impact evaluated, because the ability of research designs to detect the "little" from the midst of many

---

[14] Significance is a function of the "noise", or variance in the outcome of interest. If the change in an outcome is said to be "significant at a five per cent level", it means that, given the natural variance in that outcome, a change of such a magnitude would only be expected five per cent of the time.

competing drivers is too limited. In areas of study where the level of noise is large, this can even lead to a pessimistic conclusion that "nothing works".

9.31 If the final outcome measure is too noisy, the evaluator may still seek to detect a change in some intermediate outcome identified in the initial logic model (although the task still remains to translate the result into an estimate of final impact – a task which might be approached through reference to, for instance, theory-based evaluative models, see Chapter 5). Examining intermediate outcomes is a useful exercise in its own right, as it can help to understand the mechanism of the intervention. For example, it would be very hard to detect the effect of an advertising campaign promoting healthy eating on ultimate health outcomes, but a survey which showed some behaviour change, for instance higher consumption of fruit and vegetables in those areas subject to the campaign, might provide evidence that the campaign had had some success in communicating its message.

9.32 Even if it is not possible to detect an impact, it might still be possible to answer the question: in a best case scenario, how good might the policy benefit be, and yet have a reasonable chance of failing to be detected by the study? This could be important if it turns out that, even under such an optimistic scenario, the costs of the policy would outweigh its benefits. This can be done by deriving, from power calculation, the smallest detectable effect and then comparing the benefit that would be obtained from this impact with the cost of the policy. Notice that the two possible outcomes of this method are not symmetrical: it might find that the policy would not be value for money, even if it managed to generate the smallest detectable effect; or it might just be inconclusive, in the sense that the policy might be value for money, even at some effect size smaller than the smallest detectable.

Strategies for analysing quasi experimental data

9.33 The issues to be considered when analysing the data obtained in a study mirror those which arise at the policy design stage: identifying a comparison group and addressing selection bias. Indeed, if the policy is designed appropriately, many of the potential problems will have already been addressed. This sub-section revisits those issues from the standpoint of the tools used for analysing the data. Once again, technical details of these tools are provided in supplementary guidance.

9.34 Impact evaluation is often carried out in combination with a process evaluation. It is helpful to draw on contextual information to understand what the data truly represent. For example:

- What is meant by "treatment" in the context of the policy, and how might outcomes plausibly unfold over time as a result?

- Are the outcomes being analysed valid measures of the policy's aims? Have there been any changes in the way information is recorded that could have influenced the results?

- Was the policy implemented as intended? Are there any special cases or exceptions to be aware of?

9.35 Regression modelling plays a central role in the analysis of experimental and quasi-experimental data. Regression provides estimates of association between two or more variables, and whether that association is "significant" in the sense of being expected to exist in some wider population as opposed to just having arisen by chance in the data at hand. A regression output in isolation, however strong the "significance", is silent on the question of whether the association is causal. So the fact that there is a "significant policy effect" is not necessarily evidence that the policy caused any change to occur. (Further technical detail on regression is provided in supplementary guidance.)

9.36 Whether the analyst can go further, and infer that the policy did cause the change, depends on the context of the study. It is valid to do so if an effective random allocation scheme was used: the data are then described as "experimental". More often than not, however, allocation will not be random. What the analyst then requires is a strategy for using the observational (that is, non-experimental) data to approximate an experiment – known as an identification strategy (Box 9.F).

Box 9.F: Questions to guide an identification strategy

- Realistically, **how big** is intervention impact expected to be? Is it going to be distinguishable amid "noise"? If not, it may well not be worth proceeding any further.

- What is the (actual or projected) **comparison group**?

- Other than the policy, what else might affect the outcome?

  Is the "**what else**" effectively **random** between the treatment and comparison groups?

  So is it reasonable to believe the comparison group is **equivalent** to the treatment group (apart from the treatment, of course)?

- If it is not equivalent, it is possible to:

  - **Control** for the differences by modelling them directly?

  - Find **subsets** of the comparison and treatment groups that are more nearly equivalent (e.g. by matching)?

  - Show that the differences are **unlikely to affect** the outcome measure (e.g. from historical data, studies elsewhere)?

  And do different variants on the above give similar answers (sensitivity testing)?

  If not, what characteristics of the groups are driving the discrepancies?

9.37 The first part of the strategy involves finding one (or more) comparison groups. Ideally, the design of the policy allocation will already have provided one. Usually, the comparison group will be a group of actual subjects (people, institutions or areas). If no actual group can be identified then the comparison group might be a forecast or projection (but see paragraph 9.49).

9.38 The strategy next has to consider whether the comparison group is equivalent – that is, whether it is a plausible match for how the treatment group would have looked had it not received the treatment. For example, if the comparison group consisted of individuals who did not participate in treatment for purely administrative reasons, such as non-availability of a caseworker at the right time, it could be regarded as "as good as random" because the administrative reasons for non-participation are unrelated to the characteristics of the individuals.

9.39 Provided some basic conditions are met, control groups from RCTs, and equivalent comparison groups as defined above, provide an estimate of the counterfactual "as-is" and analysis might be relatively simple. Ideally it might only involve conducting a t-test[15] comparing the figures for the outcome of interest for the two groups; a significant difference is interpreted as evidence of a policy effect. Even in these simple cases, though, the analyst should always examine the assumptions critically in the way described in the next sub-section. If data are available on
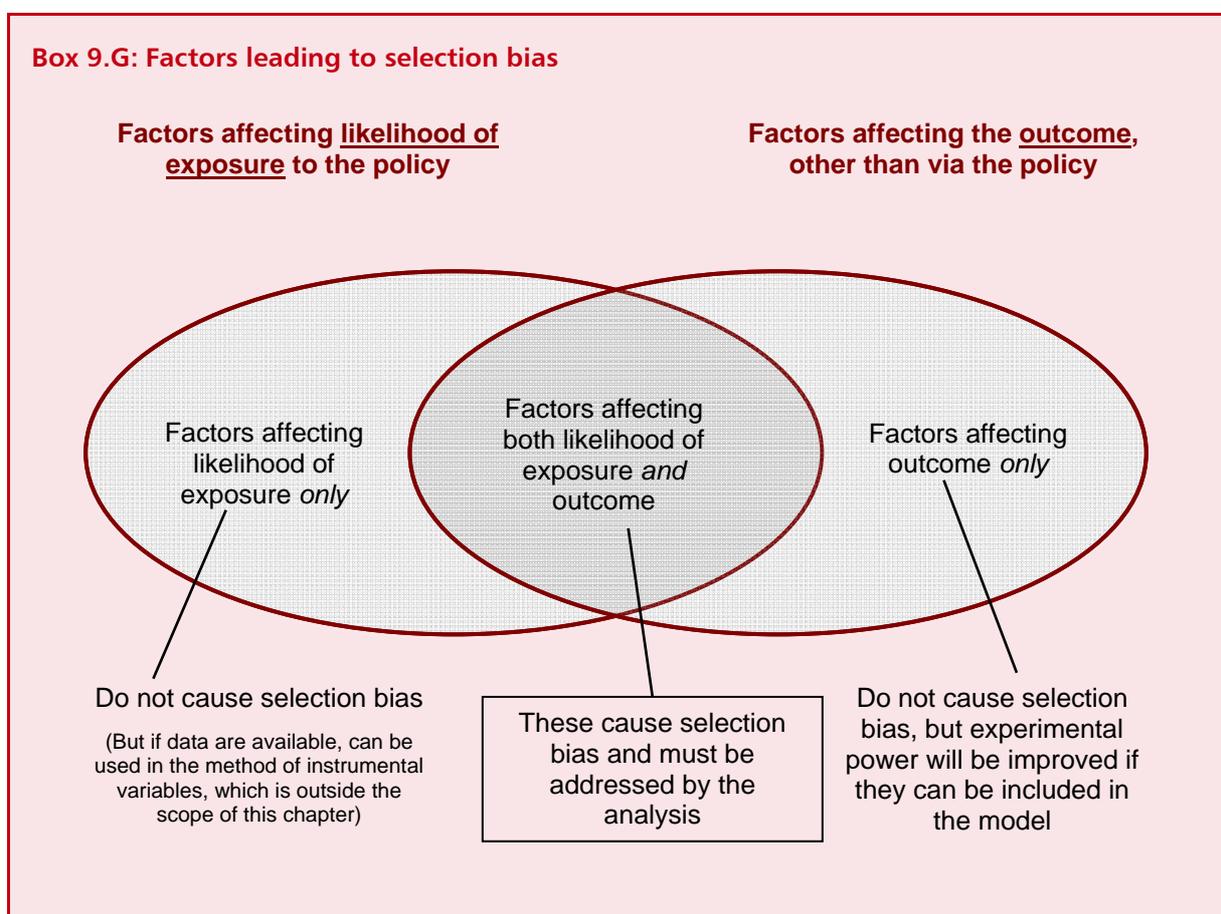
---

[15] Note that a t-test may be regarded as a special case of a regression analysis: it can always be formulated as a regression model with appropriate use of dummy variables.

additional factors thought to affect the outcome, even if they are not sources of selection bias as such, then it is worthwhile to include these additional factors in the model. This is true for all the models discussed here, not only for RCTs. Doing so improves the power of the design.

9.40 If the groups are thought to be non-equivalent, further steps must be taken to modify the model in a way that will allow any apparent policy effect to be attributed to the policy, just as it would be for a true experiment. This means overcoming selection bias as introduced in paragraph 9.14. More specifically, selection bias arises when there are factors (Box 9.G) affecting both:

1   the likelihood of an individual being exposed to the policy; and

2   the outcome measure, other than via exposure to the policy.

9.41 For example, the level of motivation of an individual to obtain a job could affect both his likelihood to enrol on a job training programme but also how likely he would be to gain employment in the absence of the programme. So a simple comparison of programme participants with non-participants would not be a valid basis on which to evaluate the impact of the programme.

Box 9.G: Factors leading to selection bias

Factors affecting <u>likelihood of exposure</u> to the policy

Factors affecting the <u>outcome</u>, other than via the policy

Factors affecting likelihood of exposure *only*

Factors affecting both likelihood of exposure *and* outcome

Factors affecting outcome *only*

Do not cause selection bias

(But if data are available, can be used in the method of instrumental variables, which is outside the scope of this chapter)

These cause selection bias and must be addressed by the analysis

Do not cause selection bias, but experimental power will be improved if they can be included in the model

9.42 Factors which affect only one out of (1) and (2) above, or which affect neither of them, do not bias the results. This points to a strategy for reducing or even eliminating selection bias. If all the factors affecting likelihood of selection are known about – as might be the case if the policy had objective selection criteria – then they can be adjusted for, in one of the ways outlined below. This will be sufficient to cover everything in the intersection region of the Venn diagram in Box 9.G, and explains why accurate knowledge of the policy allocation criteria is so valuable

to the researcher.[16] (A similar strategy could be applied if all the factors affecting the outcome were known about, but this is rarer and requires very rich data.)

9.43 The next case to consider is when policy allocation is neither fully random (as in an RCT) nor fully deterministic (as in an RDD or other objective scheme). It is this middle ground that is often encountered, because the criteria leading to exposure to the policy may not be fully known to the researcher – perhaps because they involved a subjective element, either on the part of the intervention provider or the participant. The question then to consider is whether, after adjusting for factors known to affect allocation, there are grounds for believing that whatever variation in exposure remains is "as good as random". If this is a reasonable assumption then a comparison after adjusting for these factors can proceed as in the deterministic (RDD-style) case.

Adjusting for factors affecting allocation

9.44 So, if "adjusting for" some set of factors is appropriate, how in practice is this adjustment performed? Essentially there are two strategies:

- controlling for them - the relevant factors are entered as explanatory variables in the regression model. If the policy effect remains significant in this expanded model, it is interpreted as a causal effect of the policy; or

- matching - the factors are used in a technique such as propensity score matching (PSM) to select subsets of the treated and untreated individuals that may be regarded as equivalent (in the sense defined above). A simple comparison between the matched groups might then be made, as it would be for an RCT. Box 9.I provides an example of an evaluation using propensity score matching.

9.45 When deciding which strategy to use, the first point to note is that in terms of addressing selection bias, they are equivalent. The choice therefore rests on other features of the data rather than on the assumptions being made about what drives exposure to the policy. A brief description is provided in Box 9.H.

---

[16] Within this framework, the regression discontinuity design (paragraph 9.22) may be seen as a special case of perfect objective allocation. By definition, all the factors affecting exposure are known about, because they are encapsulated in just one variable – namely the assignment score. This is what makes the RDD so effective in addressing selection bias.

Box 9.H: Control using regression

Control using regression is simple to implement, provides an estimate based on all the data, and allows the effects of relevant factors to be estimated individually. But regression models have to assume that the underlying relationship between variables has a particular shape, or "functional form" (in simple cases, just a straight line).

Departures from these assumptions turn out to be particularly problematic when the same factor strongly affects both exposure and outcome, as unfortunately, tends to be the case with quasi-experimental studies. A further issue is that the regression model will be based in part on individuals whose likelihood of participating is extremely low, and whose outcomes may bear little relationship to those of individuals who do actually participate. Matching designs have the advantage that they do not require any functional form assumption, but have their own difficulties.

For instance, depending on the success of matching they may involve discarding a significant portion of the data – especially if the targeting of the policy is such that the untreated contain few good matches for the treated. Matching can also be more complicated to implement.

The issues are technical and for a more detailed discussion of these points the reader is referred to Bryson et al.[17]

9.46 It is important to realise that both the matching and controlling approaches depend on the assumption that all sources of selection bias have been captured in the data available to the researcher. If there is "selection on unobservables", and other, unknown, factors affect the probability of treatment, then regardless of how elaborate the modelling procedure it is simply not possible to tell how much, if any, of the estimated policy effect is real, and how much is due to the unmodelled selection bias. A common example of selection on unobservables is motivation of participants in voluntary schemes, discussed earlier. A second example is personal knowledge of the candidate (for example, by a teacher, social worker, probation officer, etc.) which might affect that professional's decision to put the candidate forward for intervention. Where this is the case, an alternative approach that does not depend on identifying all the individual sources of selection bias may be stronger.

---

[17] The use of propensity score matching in the evaluation of active labour market policies, Bryson, Dorsett and Purdon, Department for Work and Pensions Working Paper No. 4 (2002). http://www.dwp.gov.uk/

Box 9.I: An example of an evaluation using propensity score matching

New Deal for Lone Parents Evaluation (Department for Work and Pensions)

New Deal for Lone Parents (NDLP) is targeted at lone parents on Income Support (IS). It tries to place job ready lone parents into paid work and to prepare lone parents not currently in the market for work for entry to the labour market. NDLP was subject to a rigorous evaluation, one component of which was to measure the counterfactual (i.e. the additional benefits of the programme). However, there were a number of challenges in meeting this aim:

- a matched area comparison was not possible because the programme was implemented in all areas of the UK;

- all members of the target group were invited to join NDLP so there was no opportunity to select a control group from individuals that had not been invited; and

- due to the relatively low take-up of NDLP, the maximum possible effect on aggregate numbers on Income Support was small, so that a time series approach to the impact assessment was not feasible.

Propensity Score Matching was chosen because it allowed a comparison sample to be drawn from lone parents who had chosen not to participate in the programme. Participants and the comparison sample were matched on their "propensity score" – the probability of participating conditional on all the factors that affect both participation and outcomes.[18] A key issue in implementing this approach was that it was well-known that motivation of individuals is linked both to participation and outcomes, and that failure to control for this would almost certainly bias the results. This was addressed by explicitly collecting baseline data on motivation/attitudes through a carefully designed survey.

A stratified sample of approximately 70,000 lone parents was selected from Income Support records using data from August and October 2000. The sample was restricted to those who, at the time of selection, had not participated in the programme. Administrative systems were used to identify those who participated and these formed the sample of "participants". The rest of the sample was categorised as non-participants, the sample of participants were matched to a comparison sample of "non-participants", using a combination of administrative and survey data, including that on attitudes.

NDLP appears to have had a large positive impact on entries into work. After six months, 43 per cent of participants had entered full-time or part-time work compared to 19 per cent of matched non-participants. This suggests that 24 per cent of lone parent participants had found work that would not otherwise have done so.

Similar effects were observed when looking at the exit rate from Income Support; NDLP appears to dramatically increase the rate at which lone parents leave benefit.

There is no evidence to suggest that NDLP jobs are not sustainable: on the whole, participants left jobs less quickly than non-participants (12 per cent of participants left work (of 16 hours or

---

[18] Many studies tend to match on whatever observable characteristics are available, whether these are the actual factors affecting participation and outcomes or not. In fact, in many situations these factors are either unobservable or simply not known, and hence should be subject to additional hypotheses.

more per week) within six months compared with 14 per cent of matched non-participants).

For more information the report is available online [19] as is a subsequent more detailed technical assessment of the results.

Making use of time trends: Difference in difference

9.47 One alternative is the method of difference in difference (DiD; or "two group pre- and post-test design"). Once again, the aim is to adjust for those factors that affect both likelihood of exposure to the policy and the outcome from the policy, and hence that might cause selection bias. But this method does so without having to know what all these individual factors are, and as such is far less data hungry. Instead, it works by comparing how trends in associated outcomes change between treated and untreated groups over a time period relevant to the intervention. While the unobserved factors might affect the outcome, if they do not affect trends in the outcome, then the trends for both groups in the absence of the policy will be the same. This is the so-called parallelism or "common trends" assumption. Any significant difference in trends is therefore interpreted as a policy effect.

9.48 The parallelism assumption should always be verified where possible, either by examining the pre-policy trends in historical time series data or from previous studies. Where the assumption does hold, DiD is a useful method that is able to address selection bias in the absence of rich information about the individuals under study. But the parallelism assumption should not be automatically assumed true, and a DiD approach would not be recommended if, for example, data are only available at two time points (before and after the implementation of the policy). Box 9.J provides an example of an evaluation using a difference in differences method.

Box 9.J: An example of a difference in difference evaluation

Multifaceted evaluation of Workplace Health Connect (Health and Safety Executive)

The Workplace Health Connect (WHC) pilot ran from February 2006 until February 2008. It was a free, no-obligation, service which aimed to provide small and medium-sized enterprises (SMEs) with advice on workplace health issues to increase the level of healthy workplaces across England and Wales.

The primary research questions were:

- whether the visit service made a net impact on the incidence and duration of occupationally related ill-health and injury; and
- what the costs, benefits, and perceived barriers to full use of the service were.

A multi-stranded methodological approach was developed to meet the objectives, which included surveys to collect data on service inputs; consider regional experiences; provide a comparator group; develop user case studies and; determine costs involved in being a WHC pilot user.

---

[19] *Evaluation of the New Deal for Lone Parents: technical report for the quantitative survey; DWP Working Age Report 146,* Phillips, Pickering, Lessof, Purdon and Hales. 2003, http://www.dwp.gov.uk/

In order to define the counterfactual for the quantitative impact study data was analysed on employers operating in regions where the WHC workplace visit service was not provided. These employers were the "comparator" group for WHC pilot users. Organisations in areas where WHC pathfinders were not in operation were selected for participation in the impact survey on the basis that they were similar (in terms of their size and sector) to those participating in the WHC pilot. Their outcomes, therefore, constitute the best available estimate of the counterfactual.

The impact survey dataset included 520 organisations within the "treatment group" and 1609 organisations from the "comparator group". Each organisation was interviewed twice, with a year between interviews, regarding a variety of health and safety outcomes.

One way of evaluating the impact of the WHC pilot would have been to look directly at the relationship between involvement in the pilot and final outcomes. This approach, however, was considered unlikely to produce robust results because in addition to improving safety using the pilot can change the way that the final outcomes are recorded.

Instead the approach taken was to analyse the relationship in two stages, looking first at the effect of the WHC pilot on intermediate outcomes and then looking at the effect of the intermediate outcomes on the final outcomes. These relationships were examined using difference-in-difference analysis. This looks at the changes in outcomes between the two survey waves, and tests whether these changes are different for the WHC pilot user and comparator groups.

In addition to the range of health and safety information gathered at the two interviews, information regarding general organisational characteristics was used to allow the analysis to control for these factors.

There was no evidence that taking part in WHC had a direct measurable effect on rates of sickness absence. There was, however, evidence that involvement with WHC lead to improvements in a range of health and safety practices. These in turn were linked to a reduction in accident rates.

The costs of the service, when the costs incurred by employers were included in the calculation, outweighed the pilot's measurable benefits.[20]

Can impact evaluation still be done when there is no physical comparison group?

9.49 A situation where there is no physical comparison group might arise if the policy was introduced everywhere simultaneously, or if there are no data available on non-participants. In this situation, the evaluator can attempt to estimate a counterfactual from a forecast or projection of the outcome measure derived from the pre-policy history, and compare it with the actual outcome. This is the basis of the **interrupted time series** (ITS) design. In practice, this design can only be used when:

- the nature of external influences is sufficiently well understood to eliminate any alternative causes; and

- the impact is sufficiently large compared with the error inherent in the forecasting procedure. In practice, only very major policy changes, that overturn a persistent historical trend, or overwhelmingly dominate sources of random fluctuation, can be

---

[20] *Workplace Health Connect Pilot: Evaluation Findings,* Institute for Employment Studies, 2009, http://www.hse.gov.uk/

detected with this method; the statistical power of an ITS is generally much lower than for designs involving a comparison group.

9.50 As a result of these restrictions, the analyst should be aware that the ITS can only be used rather rarely in public policy evaluation.

9.51 An alternative approach which is sometimes possible when there is no physical comparison group is to examine alternative outcomes which, other things being equal, have been seen to move in parallel with the one targeted by the policy. For example, a policy targeted against a particular crime type could compare outcomes for a different crime type which historically has had a similar trend; or an intervention based on cancer screening could look at outcomes for a different type of cancer. As with the ITS, the evaluator should remain alert to the possibility of reasons other than the policy why the two outcomes might have diverged.

9.52 The above discussion has not provided a comprehensive "listing" of all of the possible approaches to estimating a counterfactual. Rather, it has sought to explain the thinking behind identification strategy, and how different problems in counterfactual estimation might be addressed. The identification strategy inevitably involves making some assumptions, which in many cases can be relatively strong. Any evaluation should include an explicit acknowledgement of these assumptions, and comment on their plausibility – where it is possible to test the assumptions directly it should be done.

9.53 It is clear that each alternative approach that has been discussed has its advantages and disadvantages and it is often difficult to provide prescriptive guidance and instructions on how to go about deciding which is the best approach for a given problem situation. Judgment and common sense should drive the decision making process.

Box 9.K: When the evaluation is not (just) about individual people

So far this chapter has been couched in terms of analysing the outcomes of individual people; there are of course other types of evaluation.

In many cases, there is interest not only in the outcome of individuals, but of the units to which they belong – a good example is schools and pupils. Ideally the evaluator will have access to data at the individual pupil level, and also know the schools to which they belong. If these data are available,[21] then in many cases an appropriate approach is multi-level modelling (MLM) (more detail is provided in the supplementary guidance). This allows the analyst, in this example, to model explicitly the effects on outcomes of both school level factors and individual pupil level factors, and see which of these are more important.

In some cases however, either the data for individuals are not available, only the unit level aggregates (such as school league tables), or the outcomes are only meaningful at the unit level, such as profits data for businesses. In such cases exactly the same considerations apply in principle as for evaluation of individual outcomes. There are however likely to be differences in practice. There are likely to be fewer units in the population, making it impractical to have very large samples. The units are likely to be more diverse than individual people. And it is more likely that the intervention affects units to a differing and measurable degree (e.g. some additional source of funding for schools), which can be utilised in the evaluation.

A further degree of abstraction is when data are only available at a population level. Again, this can be either because the data are aggregated up from individual outcomes, but only the aggregates are available, or because the data are genuinely available only at population level. An example of the latter might be interest rates.

The constraints on the availability of data will guide the available analytical approaches. Where only population data are available, or where all units are affected by the intervention at the same time, time series modelling might be a viable approach. Where the degree to which units are affected is monitored and known, the marginal effect of increasing the intervention intensity can be modelled.

Thinking critically about the textbook techniques

9.54 The discussion above has stressed how the textbook research designs (e.g. DiD, PSM, RDD) may be viewed in a common framework as ways of addressing selection bias. They are not mutually exclusive. While it is true that one design may form the centrepiece of a study, it is often appropriate to combine elements of a number of different approaches. For example, the analyst can form matched groups prior to performing a DiD (and may then find that the parallelism assumption is much better satisfied than for unmatched groups). As a second example, the model for an RDD can usefully be augmented with terms for other variables thought to affect the outcome, if they are available (which will boost its power to detect the policy effect).

9.55 Once a preliminary analysis has been made the analyst should think critically about the assumptions involved and to what extent the results will remain robust should those assumptions be incorrect. This may involve triangulation with data collected through a process evaluation such as stakeholder interviews to probe whether the modelling has captured the

---

[21] In some cases, even when data have been recorded, they may not be readily available to the evaluator for a variety of reasons, such as data protection.

situation as it really is, running variants of the model under alternative assumptions, and where possible performing supporting analyses to test the assumptions directly. And, it almost goes without saying, always plot the data.

9.56 There are a number of threats to validity of research designs, some of them applying even where the design itself is very strong, as in the case of an RCT (further detail is provided in the supplementary guidance). These threats arise from the fact that the social scientist cannot usually control the experiment to the same degree as would be possible for a clinical researcher, and may be summarised under two headings:

- "Hawthorne effects" - subjects may react (either positively or negatively) to the knowledge that they are being experimented on, and in a way which affects the outcome of interest. This can occur especially if they are aware either of being granted or denied a potentially beneficial treatment. For instance, a participant who is denied access to a training course might react by seeking additional training outside of the trial. In a clinical setting this risk is mitigated by blinding or the use of a placebo, but this is almost impossible in the social policy field.

- Mis-assignment - the actual allocation and receipt of treatment may differ from what the researcher intended, because either the provider or recipient circumvented the planned design, for a variety of reasons.

9.57 Process evaluation can be valuable in determining whether and to what extent either of these has occurred.

9.58 Whenever a policy was targeted on individuals who were outliers in some way (for example, prolific offenders, low educational attainers) a common hazard for the evaluator is regression to the mean. If assignment to the policy was based on a snapshot measure shortly before it began (for instance, the number of offences in the last month, or results in a recent school test) then the selection process will to some extent capture the results of temporary fluctuations in an individual's life rather than underlying extremes. After participation, it is more likely for the extreme individuals to recover their underlying level, or "regress to the mean", than to become yet more extreme. The outcome will be seen to improve, but this will be at least partly a "natural" improvement, which, if unrecognised, might result in a misleading impression of a policy benefit.

9.59 The evaluator can check directly for regression to the mean if historical data are available, by looking for evidence that the outcome of interest has natural variability ("peaks and troughs"), and then seeing whether recruitment into a scheme appeared to occur closer to a peak. Repeating the analysis using different time baselines is a useful sensitivity test for this purpose. Some research designs, such as RCTs and RDDs, are constructed to avoid the problem making this check unnecessary, whereas others such as matching designs and DiD do not.

9.60 Examining historical time series data, where available, is valuable for descriptive purposes. It places any changes in the outcome measure that might have been the result of the policy in the context of pre-existing trends (did the trend change after the policy was introduced?) and can be used to test the parallelism assumption for DiD. Indeed, whenever a non-equivalent comparison group is used, the evaluator has considerably more confidence that post-policy changes were caused by the policy if the comparison and treatment groups have tracked one another for a long historical period. A useful trick when visually examining the data is to index the time series to a common baseline.

9.61 Another judgement the evaluator will wish to make is whether a "matched" comparison group really is matched. With regard to observed characteristics, this may be done by comparing distributions between the two groups. This check should be done even for RCTs, especially when numbers are small, as randomisation does not always provide balanced samples – that is,

samples which are similar in terms of the characteristics likely to affect the outcome. With regard to unobserved characteristics, careful consideration based on subject area knowledge will be needed to assess possible non-equivalence.

9.62 A particular case where a "matched" comparison group may fail is when policy allocation was in fact rigorously targeted, but the evaluator does not have access to all the information on which the targeting was based, perhaps for one of the reasons mentioned in paragraph 9.56. In this case, the presence of a reasonably sized region of common support[22] should be regarded with the utmost suspicion: it is virtually a sure sign that the selection bias has not been adequately captured, because in a deterministic selection process there should be no common support at all (just as there would not be for an RDD). A DiD analysis on the "matched" groups might provide a remedy (assuming the historical data exist to permit it), since it acknowledges the non-equivalence of the two groups.

9.63 As with any statistical study, the evaluator should beware of embarking on "fishing expeditions or data mining", especially when many variants of a model are being fitted. If different variants give different conclusions it is vital to be clear about how the assumptions differ and the robustness or otherwise of the model to changing them. A useful technique is to hold back a portion of the data during an initial phase of analysis and then check that these data give consistent results.

"Constrained designs"

9.64 Much of this chapter has been concerned with the design and analysis of studies when the policy has been designed so as to provide a comparison group. However, an analyst may be asked to evaluate a policy that is not amenable to these approaches, for example, if on practical grounds none of the desired policy allocation methods was possible, or if data are not available or of insufficient quality, or the policy has already been implemented and the opportunity to put a research design in place was missed.

Natural experiments and instrumental variables

9.65 A solution may present itself if it is possible to carry out any of the approaches in this chapter in retrospect. The influence of random shocks or administrative anomalies on policy allocation can sometimes create a so-called "natural experiment", in which comparisons with a naturally occurring comparison group can be made even though none was present by design. Essentially the same theory and analysis considerations then carry through. A more general case is where a so-called instrumental variable can be identified – an external factor which influences the likelihood of being exposed to a policy, and which does not in itself affect outcomes. This can be a very useful way of overcoming selection bias. It is often difficult, however, to find a suitable instrument, and very rare to identify one in advance, so it is not common to use this as part of a planned evaluation strategy. More information on this approach is given in the supplementary guidance.

"Before and after" studies

9.66 Sometimes the level of evidence available falls far short of what would generally be regarded as a true impact evaluation. A common example is the single group pre-and post-test design, or simply "before and after" design, in which an outcome is measured before and after intervention takes place but there is no comparison group. This only really has any credibility when the system being studied is so simple that the policy is the only thing that could reasonably be expected to influence the result. Unfortunately, real social systems are seldom that

---

[22] The "common support" consists of those members of the treatment and comparison groups who can be matched to each other. It is discussed in more detail in the supplementary guidance.

simple. Unless there is a strong justification for ruling out influences other than the policy (not simply a lack of obvious alternative explanations), this design should not be reported as an impact evaluation. The supplementary guidance provides detail on the large number of threats to validity with this design.

Use of process evaluation information

9.67 This chapter has already highlighted the benefits of combined evaluations where process studies, which study the implementation and delivery of a policy or intervention often using qualitative methods, (Chapter 8), are integrated with impact evaluation. This is particularly important when quantitative measures of impact are weak, or not available at all. If as above there is no comparison group, or worse still not even an outcome measure is available, then the researcher may be able to draw upon the findings of a process study, action research or case studies. By their nature these types of study do not allow a quantitative measurement of impact, but they may be able to capture a direction of change. Front line staff directly involved in the delivery of the intervention will have a good feel for whether or not it is effective, and why. Care must be taken, however, that the evidence captured reflects the achievement of the wider aims of the policy, and is able to look beyond the immediately perceived impact by the interviewees.

Reporting of evaluation results

9.68 Whichever approach is used, the evaluation report should be worded to give an accurate and objective reflection of the strength of the evidence. If there remain significant doubts as to the strength of the counterfactual estimate (or if it could not be estimated at all) then the evaluator should avoid using the term "impact" or any other wording that would imply attribution of the outcome to the policy. Only if the evidence points decidedly towards a causal effect of the policy should it be reported in these terms. As usual, any appropriate caveats with regard to the assumptions made and the strength of the available evidence should appear alongside the conclusion.

> The guidance in this section of the Magenta Book has been revised since the previous edition to clarify that weak designs, where there is no compelling reason to ascribe the outcome to the policy or to eliminate other potential causes, should in general not be reported as impact evaluations.

9.69 As an example of appropriate reporting, the results of a successful (fictitious) impact evaluation might be stated as follows.

9.70 "The results of the ABC pilot imply that the proportion of pupils achieving five grades A-C at GCSE was increased by 0.7 per cent as a result of the ABC programme. This is after taking into account known differences between participating and non-participating schools, though there remains a possibility that some other differences between the schools could have contributed."

9.71 If a true impact evaluation was not possible, the evaluator should avoid wording like the following:

9.72 "In the year following the nationwide rollout of the XYZ policy, the proportion of pupils achieving five grades A-C at GCSE rose by 1.2 per cent. It is not possible to say for sure whether this was the result of the policy, but the results are encouraging."

9.73 This is bad reporting. There is too much risk of the first sentence being taken out of context. Despite the "caveat", the report seems to want to imply that the XYZ policy caused the improvement. To a casual reader the strength of evidence might seem to be similar for both the

ABC pilot and the XYZ policy, when in reality the former is pretty robust but the latter is paper-thin. It is true enough that an increase in attainment is better than a decrease, but in order to regard it as "encouraging" (from the point of view of the XYZ policy) we would require a much wider appreciation in the context of other drivers of change and previous trends.

9.74 If the previous example could be backed up by some qualitative evidence, a more appropriate form of words might be:

9.75 "Although in the year following the nationwide rollout of XYZ policy the proportion of pupils achieving five grades A-C at GCSE rose by 1.2 per cent, this welcome rise was not necessarily caused by the policy. For such a claim to be made with confidence would require an appropriate evaluation that controls for other factors. However, interviews with teachers suggested that the policy had filled a genuine gap for struggling pupils who in previous years might have fallen through the net. It is therefore reasonable to suppose that it has contributed to the 1.2 per cent increase in proportion of grades A-C in the year since it was introduced."

9.76 To conclude, this chapter has described how the evaluator can go beyond merely stating what happened, and report something much more relevant to the policy maker: namely, whether the policy caused it to happen. The rationale for doing the extra work required is that it answers the impact evaluation question, whereas descriptive statistics alone do not. The two types of evidence – descriptions of the situation on the one hand, and impact evaluations on the other – say very different things and need to be reported in correspondingly different ways. The one must not be misrepresented as the other.

10 Drawing together and reporting evaluation evidence

<div style="border:1px solid red">

Key points

- How the findings of an evaluation will be used and disseminated must be considered at the planning stage of the evaluation.

- A strategy for synthesising evaluation evidence should be agreed in advance, to avoid any possible accusations of picking the results which best support a particular viewpoint.

- Evaluation results should be set in the context of other knowledge about the intervention and/or the context in which it was delivered.

- A thorough evaluation can be time-consuming and/or expensive. It is important to get the maximum value from the investment, for example by ensuring that results can and do feed into important decision-making processes such as spending reviews.

- Decisions about future policy will not be made solely on the basis of a single evaluation.

</div>

Introduction

10.1 This section provides guidance on how to draw together qualitative and quantitative evidence from a programme of evaluations and set the findings in a broader context. The section also considers the implications of this for initial evaluation planning and discuses the presentation and dissemination of findings to ensure they impact on future decision-making and rolling-out/scaling-up of the policy where appropriate.

How evaluation evidence may be used

10.2 Evaluation evidence can be used to inform a range of different types of decisions, such as:

- immediate decisions about policy options; for example whether to roll-out a pilot as a national or local programme;

- longer term decisions about the policy/programme; for example informing Spending Reviews and the future scale of investment;

- how the programme/policy could, or should, be improved; for example if the evaluation identifies major flaws; and

- how future policies should be designed and implemented.

Drawing together the evaluation evidence

10.3 An important task in all evaluations is to bring together the evidence collected from different parts of the evaluation to present a complete account. What are the answers to the

original research questions? Do the results support each other, or are there apparent contradictions? In a small-scale evaluation this may be a fairly straightforward task, but in others, with many separate studies (such as process and impact evaluations) carried out over a number of years, it can be substantial. It is important to ensure that sufficient time and resource is allocated for this part of the evaluation.

10.4 When drawing together quantitative and qualitative evaluation evidence it is important to consider whether answers to different questions are consistent. A process evaluation might, for example, find that a policy was only weakly implemented; yet the impact study shows that it still had a significant effect. The different parts of the evaluation will need to be used to examine the original logic model (see Chapter 5 for further detail on the use of logic models).

10.5 Ideally, all the steps in the model are found to work as anticipated: a programme is implemented as intended; participants change their behaviour as predicted; and the desired outcomes are observed. Where this occurs, the overall consistency of the various evaluation findings increases our confidence in them. However, there may be occasions where some steps cannot be fully validated, for example, all the processes are seen to have worked as expected, but there is only weak evidence of overall impact. In such a case, confirming the earlier steps in the logic model will lead to increased confidence that the observed impacts are genuine.

10.6 But in some cases this does not happen, and the logic model breaks down. This can occur at a relatively early stage in the model. For example, suppose that the evaluation of a training programme for unemployed people finds that there is no significant impact, and that a large proportion of participants drop out before completing the training. We can then look for evidence as to why this happened using other evaluation evidence, for example through qualitative studies of participants, exploring why they did or did not complete a course, or through more detailed analysis of quantitative data to identify what factors are statistically associated with completing a training course.

10.7 Sometimes the break in the logic model can be at a later stage: a policy is fully implemented as intended but does not have the desired impact. For example, a programme is designed to help move unemployed people into work by encouraging them to search more actively for jobs, based on previous evidence suggesting that this will result in faster movement into work. The evaluation shows that people participate in the programme, and intensify their job search but that there is no impact on employment. Again, other parts of the evaluation may suggest explanations for this, for example there may be evidence that the current state of the labour market reduces the effectiveness; or that the programme only works for certain sub-groups of individuals.

10.8 It is extremely important to note that these conclusions are not robust findings in their own right, but are new hypotheses which will need further testing to verify them. (A good treatment of the iterative process of refining hypotheses in this way is given in Pawson and Tilley's book on realistic evaluation.)[1]

10.9 It is useful to capture and document these emerging hypotheses as changes to, or refinements of, the original logic model, being careful to distinguish between those parts which are clearly supported by evidence, and those which are for further testing.

10.10 It is highly advisable to set down in advance the intended strategy for reconciling different estimates of impact. For example the Pathways to Work evaluation[2] collected data on a cohort of those joining the pilots early in their operation in addition to a cohort that joined after the

---

[1] *Realistic Evaluation*. Pawson and Tilley.1997 – see Chapter 5 in particular

[2] *Pathways to Work for new and repeat incapacity benefits claimants: Evaluation synthesis report, Research Report No 525* National Institute of Economic and Social Research on behalf of the Department for Work and Pensions, 2008. http://www.dwp.gov.uk/

programme had been operating for six months, by which time it was expected that initial teething troubles would have been addressed. The intention was always explicit to use the results from the later, "preferred", cohort. It is important to set this out early on because otherwise it can be difficult to avoid accusations of choosing evidence to support a prior viewpoint.

10.11 There are no hard and fast rules for this process of drawing data together and many analysts will already have experience of synthesising data. For those wishing to learn more there are textbooks on the topic, for example Cooper and Hedges (1994).[3] It is worth noting that there are separate considerations for quantitative and qualitative data.

Synthesising quantitative data

10.12 One of the most common quantitative synthesis tasks is to reconcile a number of different assessments of impact which may be based on different:

- data sources – for example survey and administrative data;

- groups of affected individuals – for example the first and final waves of recipients to receive an intervention, as in the evaluation of the impact of Pathways to Work; or

- statistical approaches and assumptions - Chapter 9 explained how the validity of the impact assessments depends on key assumptions.

10.13 It is highly unlikely that all the estimates will have equal validity meaning that a statistical combination of them to give an overall best estimate will not be possible. There are two types of validity to consider here: internal and external, as discussed in Box 10.A.

Box 10.A: Considerations of internal and external validity

Internal validity (as discussed in Chapter 9 paragraph 9.14) refers to whether the results are a true reflection of the impact on the individuals being studied. In the case of a pilot study for example, are the estimates a true reflection of the impact on the individuals in the particular areas involved in the pilot during the lifetime of the evaluation? All statistical approaches to impact estimation depend on assumptions. Where different statistical approaches have been followed, it will almost always be because it was not possible to be certain in advance whether the necessary assumptions hold. Where possible, formal tests of the validity of the assumptions should be carried out (for example, testing the common trends or parallelism assumption in a difference-in-difference design. See Chapter 9 for a more detailed discussion).

External validity refers to whether the impact estimated for those directly studied can be extrapolated / generalised to others. For example, as in the Pathways to Work example, the impact of a programme on the first group to go through it is likely to be a poor guide to its effectiveness, due to teething problems. A better guide is likely to be the impact on those who experience it after it has bedded in. More discussion of potential threats to external validity is given in paragraph 10.28.

10.14 A different type of consideration might be which data source is closest to measuring the relevant outcomes. Administrative data would normally be more accurate than self-reported data where something very specific and objective is being measured. For example, it is well known that survey responses about which welfare benefits claimants receive are not fully

---

[3] *The Handbook of Research Synthesis,* Cooper and Hedges (Eds), 1994, New York: Russell Sage Foundation

reliable. Administrative data sources, in many cases, will have extremely low sampling error, giving far greater precision than is possible with surveys. But very often administrative data and surveys are measuring different things, or there are known limitations about one of the sources. For example, administrative data can provide information about numbers of recorded crimes, but only surveys can provide data on the fear of crime. Chapter 7 discusses surveys and administrative data in more detail.

10.15 On a related point, there is also the question of which results answer most closely the question at hand, which in turn depends on the decision being made. As explained in Table 10.A, impacts can be either average or marginal. Where the decision being made is whether or not to continue with a policy, or to implement a pilot, it is appropriate to use average treatment effects. But where the question is whether to expand or contract a programme, marginal effects are more important. As previously noted, which of these is available is likely to be dictated largely by circumstances rather than by choice. Where it is necessary to make decisions based on average effects when marginal effects would be more appropriate, or vice versa, it may be possible to explore the heterogeneity of treatment effects, either quantitatively (for example, looking at impacts for sub-groups) or qualitatively. The need for this should be considered at the planning stage.

10.16 In some cases, it may be clear that one set of estimates is more likely to be valid than others, and is therefore the appropriate one to use. In other cases, sampling error may explain the differences allowing the findings to be combined arithmetically. There may be occasions however where, despite best efforts, it may not be possible to fully reconcile the different studies, in which case it may be appropriate to report the impact as a range rather than as an exact figure.

Table 10.A: Types of impact estimates

Types of impact estimates		
Intention to treat (ITT)	The impact of the policy on the target group. For example, for a training programme for jobseekers, the net impact on all those eligible, whether they participated or not.	Where participation is voluntary, estimating the impact on the Intention To Treat group avoids most of the problems of selection bias. But where the proportion participating is small, the impact is small and can be very hard to detect.
Treatment on the Treated	The net impact on those who were actually affected by the intervention – for example, those who took part in a training programme.	It will be much easier to detect with small participation rates, but depending on how participants are selected it may be difficult to account for bias.
Which of these is estimated is more likely to depend on which impact evaluation methods are feasible than on which is more desirable. Note that as long as it is known who is treated and who is not, and that it is reasonable to assume that there is no impact on the non-treated, it is straightforward to calculate one from the other.		
For either of these, there are two types of estimate:		

Average Treatment Effect	The average net impact across all those treated, or who were intended to be treated.	This is the most common, and is the preferred estimator for cost-benefit analysis in particular, and for overall decisions about whether to implement a policy. It is less suitable where the decision is about the expansion or contraction of a policy.
Marginal Treatment Effect, or Local Average Treatment Effect	The impact on those who in some sense are on the margins of participation.	An example of this is in a general Regression Discontinuity Design where the impact estimated is for those whose scores are on the borderline of eligibility. This is the estimator needed to inform decisions about expansion/contraction (in this example, changing the threshold score) but further assumptions are needed to produce an overall cost-benefit analysis.

In most cases, whether the impact estimate is marginal or overall average will depend on the available evaluation methods rather than on what is desired.

Synthesising qualitative data

10.17 Similar issues to those raised above are also relevant when synthesising qualitative findings, such as those that might be collected through a process evaluation. Process evaluations are often designed to capture the experiences of different people, areas, or institutions for example, subject to a policy, so that these differences (and similarities) can provide powerful information about its implementation and an explanation for observed impacts. However, it is important to be confident that any differences observed through qualitative research (either in the same or in separate studies) are due to actual differences in the people, groups or areas being studied rather than being the result of shortcomings in the research itself. This means that it is essential that qualitative research is designed, conducted and analysed in a way that allows confidence in the robustness of its findings. Process evaluation, action research and case studies are discussed in further detail in Chapter 8.

10.18 There are a range of approaches to assessing the quality of qualitative research ranging from using criteria similar to that used to assess quantitative data (external and internal reliability and validity) to ones specifically designed for qualitative data (credibility, dependability, confirmability and authenticity). There is a useful discussion of this in Bryman (2001)[4]. Key questions to consider when reviewing the quality of qualitative research are provided in Box 10.B.

---

[4] *Social Research Methods.* Bryman. 2001. Oxford: Oxford University Press

10.19 Once a judgement has been made that findings are valid then data from qualitative research can be presented, highlighting the different sources of this data, and signposting any differences and/ or similarities between different research participants and areas. These similarities and differences are key issues in comprehending how a policy was implemented and delivered and so the more richly they can be described and explained, the better the policy can be understood (and compared to previous research on the policy or similar policies). This doesn't mean that the findings should necessarily be reported in a long and detailed manner. The key issue will be to answer the original research questions highlighting the different or similar experiences of the policy and explaining why these might have occurred. Where it is useful to provide particularly detailed accounts these can be annexed or presented in a technical report.

Setting the evaluation results in a broader context

10.20 When considering the evaluation findings it is vital not to neglect the broader context. In addition to analysing the findings from different parts of the evaluation for reinforcement or contradiction, it is important to review the broader research evidence, including related evaluation studies and any other relevant literature. Evaluation findings will be strengthened when they are in line with earlier research. In contrast, differing findings can be explored further in order to seek explanations, thereby making valuable extensions to existing knowledge.

10.21 When seeking to understand why there are differences, it is important to look at the context in which evidence is gathered. For example, the findings may be from research undertaken abroad, such as the USA, and differences in context between the two countries need to be taken into account. For example when looking at issues around health and disability the differences in the healthcare infrastructure might be relevant. While research into criminal justice would need to take into account the differences in sentencing policy.

10.22 Another major difference in context might be temporal differences between the previous research and the current evaluation. For example, the economy may be at a different stage of the business cycle or there may have been legislative or societal changes, such as the increase in access to the Internet, which could explain the differences in the findings observed. If two evaluations are separated in time, the context in which they are carried out, for example economic, social, political, legal or technical, will inevitably have changed.

10.23 It may sometimes seem that the results of an evaluation are almost identical to previous work, questioning its value. But the conditions under which the evaluation takes place will always be different. A labour market programme, for example, might have been found to be effective at a time when the economy was expanding; finding that it is still effective when the economy is in recession would be important learning. Box 10.C provides a list of questions to consider when reviewing the broader research evidence

---

Box 10.C: Questions to consider when reviewing the broader research evidence

- What was the economic, social, political, legal or technical context within which the research was undertaken?

- Was the research undertaken in the UK? If not, are there any relevant differences in context between the UK and the country in which the research was undertaken?

- If undertaken in the UK are the geographical areas comparable in nature? For example urbanisation, levels of deprivation, etc.

- How long ago was the research undertaken? Have there been any relevant changes in context since the study was undertaken?

- Were the studies conducted at the same time of year? Could there have been any seasonal or temporal differences?

10.24 Quantitative estimates of a policy's impact may lend themselves to meta-analysis.[5] This can be used either to get more precise estimates of a policy's impact using findings from a number of different evaluations than are available from a single study; or to understand what factors are associated with varying scales of impact. Suppose there is a policy which is expected to have different impacts at different stages of the economic cycle. In principle a statistical model can be built incorporating the impact estimates at different stages of the economic cycle, and a suitable measure of the state of the economy, to test and quantify the relationship.[6]

10.25 Even where such formal meta-analysis is not possible, either because there are not enough comparable studies, or because the evidence is qualitative rather than quantitative, it is important to look at the degree of consistency between the evaluation and previous evidence (which should not be limited to previous evaluations). There are clear parallels to the previous section on synthesising evidence within an evaluation. Where the new evidence is at odds with previous studies, it may be possible to develop hypotheses about which factors influence the results. And when the new evidence is weak, it is more likely to be given credence if it is broadly consistent with earlier findings. Where it is not, it may well be an anomalous result.

Future decisions and roll-out; scaling-up

10.26 Evaluations are often undertaken of pilot programmes,[7] this section focuses on the decision whether or not to move from a pilot to a fully implemented national policy or programme.

---

[5] Meta-analysis is the process of combining statistical information from separate studies, using a range of statistical techniques.

[6] An example of meta-analysis in the criminal justice field, which explores which programme features are associated with greater effectiveness, is: *A rapid evidence assessment of the impact of mentoring on re-offending: a summary,* Home Office Online Report 11/07, Jolliffe and Farringdon, 2007, http://homeoffice.gov.uk/. An example from the labour market field is: *When welfare-to-work programs seem to work well: explaining why Riverside and Portland shine so brightly*; Greenberg et al ;Industrial and Labor Relations Review, vol 59, no.1, pp34-50

[7] In this context a "pilot" refers to a programme or policy introduced on a limited basis – for example limited in time or geographical scope with the express purpose of producing evaluation evidence to inform a decision on whether or not to proceed to full implementation. For a good discussion of

10.27 For an evaluation to have maximum impact on this decision, it is important to be certain that the results are internally valid and are an accurate reflection of the experience of those who have been affected by the pilot. Furthermore, deciding whether to move to full implementation also requires external validity, or certainty that the pilot findings can be extrapolated to estimate what would happen in a full implementation. This has a number of considerations, often referred to as "threats" to external validity which are summarised below, with examples in Box 10.D.

10.28 Reasons why results may not be generalisable, or threaten external validity, include:

- that pilot data are not representative of the wider population;

- the state of the economy at the time of the evaluation;

- what other policies and programmes were operating at the same time and in the same areas as a pilot;

- spillover effects - where for example a policy implemented in one area has effects in neighbouring areas (which may be positive or negative);

- substitution and displacement effects - where there may be positive impacts on those directly affected by a policy or programme, but negative effects on others;

- general equilibrium effects - the overall impact on outcomes taking into account any indirect or secondary effects;

- scalability - whether sufficient resources exist to implement a policy more widely. This is wider than just finances, for example a health intervention may require input from doctors who may be in short supply; and

- what are known as Hawthorne effects - where an initial pilot is successful but largely as a result of increased oversight.

10.29 To an extent it is possible to mitigate these risks by careful planning of the evaluation.

One potential threat is that those affected by a pilot are not representative of the wider population. For example, if a policy is only piloted in parts of London, it would be unwise to assume that the observed effects would be the same in other parts of the country. A well-designed pilot study would address this by including a variety of different types of area. Even so, it is unlikely to be an exact representation of the whole population. Where it is possible to quantify how the pilot areas differ from the country as a whole, it may be possible to correct for this bias. This can be particularly valuable if the choice of pilot areas (or participants) is constrained, for example, if there is a greater than average representation of urban areas in the pilot.

As an example, suppose that there are 100 areas in the country, of which 20 are urban and 80 rural. A pilot programme is run in four urban and four rural areas. Weighting the results of urban areas by 0.2 and those for rural areas by 0.8 will ensure that the overall results are, at least in this respect, balanced. This can readily be extended to two or three factors. In reality, there are likely to be more factors than this, and achieving an exact balance will not always be possible. In such cases, it may be possible to estimate overall effects in a regression framework.

A more difficult case to deal with is where the pilot areas (or people, or units) are self-selecting, for example, if local authorities were asked to volunteer to participate. In such cases, the generalisability of the pilot findings to areas that are compelled to participate in a later implementation stage cannot be assumed. This is because the characteristics and contexts of the local authorities that volunteered may have contributed to them volunteering in the first place and to the impacts observed, these factors may be different in the authorities taking part in the later implementation and may affect the impacts.

10.30 It is important to recognise that a policy evaluation that shows a positive impact and good value for money does not mean that it was an appropriate policy, similar or better gains may have been realised by alternative policies that have not been evaluated. Decision making is also a balancing of risks. Proceeding with a policy for which the evidence is weak risks wasting the resources necessary to implement it. But not proceeding in such a case risks forgoing genuine gains which would have been made if in fact the programme were effective. In each case, the strength of the evidence on impact needs to be considered alongside the potential gains from an effective programme, the potential losses from an ineffective one, and the desirability or otherwise of any unintended consequences.

Implications for evaluation planning

10.31 The extent to which the evaluation findings can be synthesised will depend on how well the evaluation has been designed and planned. Some key points to remember include:

- where an evaluation includes more than one study (for example a separate process and impact evaluation), they should be designed to complement each other;

- research questions must be clearly identified; and

- it is essential to work from an overall set of research questions rather than drawing up separate ones for each study. Consistency with external sources and previous relevant research is desirable, to enable comparison (this is discussed in more detail in Chapter 7).

10.32 Probably the most important point, though, is the necessity of planning the evaluation carefully to ensure that it can provide the necessary evidence to answer the research questions. See Chapter 5 for further guidance on planning an evaluation.

Reporting and disseminating findings

10.33 However carefully planned and meticulously conducted the evaluation, if the findings are not understood and used correctly, the research will not meet its objectives. There are some key points to take into account when reporting and publishing research and evaluation findings.

10.34 Reporting an evaluation means more than writing a final report. It is important to ensure that feedback is provided to all the evaluation stakeholders, and that findings feed into new policy development and appraisal.

10.35 Notwithstanding the range of activities that should be considered in disseminating findings, the evaluation report is a key output and its effectiveness will depend on the brevity and clarity with which key conclusions and messages are conveyed. The aim of the reporting process throughout a project is to ensure the evaluation commissioners, partners and stakeholders are consulted about research methods, progress and results on an agreed basis. Regular interaction between the evaluators and the commissioning partners maintains the focus of the evaluation and teases out any problems with data collection or team dynamics as soon as they arise.

10.36 Opportunity to reflect on the findings as soon as possible helps the stakeholders to prepare for the conclusions and recommendations, and makes hard messages easier to respond to before the final report becomes public. Subject to commissioning partners' views, allowance should be made for comparison of the evaluation results with other relevant evidence, wider dissemination of the results, and consideration of their implications for policy design and delivery.

10.37 Useful guidance is provided by Vaughan and Buss[8] on how to report social research findings to busy policy makers. They point out that many policy makers are able to read and understand complicated analysis, but most do not have the time. Consequently, many will want to be given a flavour of the complexities of the analysis but without getting lost in details. Other policy makers may not have the technical background and will want a simpler presentation. So there is a delicate balance between keeping the respect and interest of the more technical while not losing the less technical.

10.38 Of course, what the evaluation commissioners and other key stakeholders want to see and how they want to see it must determine the form and content of the report. Nevertheless, there are some simple tips suggested by Vaughan and Buss that are likely to be helpful whatever the form of the report; they are set out in Box 10.E.

---

[8] *Communicating Social Science Research to Policy makers,* Vaughan and Buss, 1998, Applied Social Science Research Methods Series No. 48. Sage Publications.

10.39 As discussed above, a useful first step is to report how the new evaluation findings compare with previous knowledge, particularly where there are clear consistencies or inconsistencies. New hypotheses may be required to explain the latter. It is useful to highlight research questions that emanate from the evaluation to inform future planners of research programmes and evaluations.

10.40 It is also important to thoroughly document the research methodology, commonly as part of a separate technical report rather than in the main report. (It is essential that the information remains available, even after all those working on a project have moved on). This should include research tools, such as questionnaires and topic guides used for qualitative/quantitative studies, as well as associated documentation, such as introductory letters and explanatory leaflets. Steps taken to process and analyse the data should be fully recorded, including:

- data cleaning or imputation of missing values;

- weighting for non-response;

- how a final statistical model was selected; and

- how standard errors were calculated.

10.41 Where possible, the source data should be archived to allow subsequent secondary analysis. Anonymised data can be deposited with the Economic and Social Data Service http://www.esds.ac.uk/, although this is more common for quantitative data. It may also be necessary to retain the identifying details separately so that survey respondents can be re-contacted for further research, or to allow linking with other data sets. Where this is the case respondents will need to provide informed consent (this is discussed in more detail in Chapter 7).

10.42 In summary it is vital to think about the dissemination of the results at the time of planning the evaluation, including how they will be used, shared and built upon.

Publication

10.43 Finally, there is the issue of publication and the form that this should take. Departments, devolved administrations, and their agencies will have specific protocols and procedures for this which should be followed and which can be discussed, as needed, with the relevant Head of Profession/Senior Analyst.

10.44 However, in general terms the case for publishing the results of evaluations and information about methodological approaches and research instruments is three-fold:

- it is an integral part of public accountability;

- it helps to improve the credibility of findings by opening them up for wider peer review (NB the importance therefore of including a clear account of the context in which the research was planned and carried out as well as details of the research methodology); and

- it contributes towards a learning legacy that transcends the passage of time and people. Credibility is also served where detailed evaluation reports are produced and made publicly available, where their findings are presented and discussed at academic and research gatherings, and they find their way into public datasets.

10.45 In order to maximise the impact of the evaluation research a dissemination strategy should also be considered. It will not be practical to have tailored outputs for each possible audience so it will be necessary to prioritise, taking into account factors such as who funded the work; who the stakeholders are; and who is in a position to react to the findings.

10.46 One particular avenue for dissemination that is worth considering is publication in a recognised journal. There can be benefits for the researcher and the commissioning government department, including:

- greater credibility for the research;

- wider dissemination of the results;

- exposure to peer review before publication (although as noted in Chapter 4, peer review can be undertaken without a journal publication); and

- critical scrutiny after publication.

10.47 For these reasons, it is usually worth allowing and even encouraging such publication.