Part A

This part of the Magenta Book is written for policy makers. Chapter 1 explains the benefits of undertaking good evaluations, and some of the difficulties that might be encountered if evaluations are not undertaken or are undertaken poorly. Chapter 2 explores the types of questions that evaluations can answer and provides an overview of the different types of evaluation that can answer these questions. It also introduces some of the issues which affect how well a policy can be evaluated and the implications this might have for the type and design of evaluation which is most appropriate. Chapter 3 considers the features of the policy itself that can affect how well the policy's impacts can be evaluated, and discusses minor adjustments which can be made to improve the chances of a good quality evaluation. Finally, Chapter 4 considers some of the practical aspects of planning an evaluation.

Chapter 1: Key issues in policy evaluation

Chapter 2: Identifying the right evaluation for the policy

Chapter 3: Building impact evaluation into policy design

Chapter 4: What practical issues need to be taken into account when designing an evaluation?

1 Key issues in policy evaluation

Key points

- Evaluation is an objective process of understanding how a policy or other intervention was implemented, what effects it had, for whom, how and why.
- Evaluations need to be tailored to the type of policy being considered, and the types of questions it is hoped to answer. The earlier an evaluation is considered in the policy development cycle, the more likely it will be that the most appropriate type of evaluation can be identified and adopted.
- Good-quality evaluations generate reliable results which can be used and quoted with confidence. They enable policies to be improved, or can justify reinvestment or resource savings. They can show whether or not policies are delivering as planned and resources being effectively used.
- Good-quality evaluations can play important roles in setting and delivering on government priorities and objectives, demonstrating accountability, and providing defensible evidence to independent scrutiny processes. They also contribute valuable knowledge to the policy evidence base, feeding into future policy development and occupying a crucial role in the policy cycle.
- Not evaluating, or evaluating poorly, will mean that policy makers will not be able to provide meaningful evidence in support of any claims they might wish to make about a policy's effectiveness. Any such claims will be effectively unfounded.

Introduction

1.1 This chapter provides an introduction to evaluation and outlines where it fits in the policy cycle. It explains what evaluation is, why it is important to evaluate and what the costs are of not evaluating, or of evaluating poorly.

What is evaluation and what benefits can it bring?

1.2 The primary focus of the Magenta Book is on policy evaluation[1] which examines how a policy or other intervention was designed and carried out and with what results.

1.3 Therefore, the focus is on the actual practice and experience of the policy and observations on what actually happened following implementation (rather than what was expected or intended, for instance, which is the topic of appraisal).

Evaluation can employ a variety of analytical methods to gather and assess information, and the choice of methods employed in any particular instance will depend on a wide range of factors which are the subject of the remainder of this book. In turn, this choice will affect what

[1] The Magenta Book generally uses the term 'policy evaluation' to refer to evaluations covering projects, policies and programmes. How evaluations differ across these various types of intervention is discussed in Chapter 2.

questions the evaluation might be able to answer and how strongly its conclusions can be relied upon. However, the focus on actual experience of a policy means that evaluation as described here is an impartial process which asks objective questions such as:

- What were the impacts of the policy?
- How was the policy delivered? and;
- Did the policy generate value for money?

1.4 Even when an evaluation asks a question on a subjective topic (such as stakeholder perceptions of effectiveness), it will seek to answer it in an objective way, such as:

- How successful did stakeholders think the policy was in achieving its objectives?
- Did the policy succeed in improving the public's perceptions of the problem?

1.5 In practice, of course, questions will be more complex and specific than this, and will often include consideration of how different features of the policy affected the way it performed and delivered, and how its outcomes varied across those it impacted upon: what worked for whom in what circumstances. The types of questions which different types of evaluation can answer are the subject of Chapter 2. Good evaluation, as described in this book, is an objective process, therefore the answers it provides will give an unbiased assessment of a policy's performance. For this reason, evaluation results might be challenging in real terms and from a presentational perspective.

1.6 However, good evaluations should always provide information which could enable less effective policies to be improved, support the reinvestment of resources in other activities, or simply save money. More generally, evaluations can generate valuable information and contribute to a wide range of initiatives and objectives. For instance good evaluation can:

- provide a sound scientific basis for policy making, by providing reliable understanding of which interventions work and are effective. An understanding of how and why policies work can also be used to inform the development of new policies, and to improve the effectiveness and reduce the burden of existing ones;
- underpin practical resourcing and policy making exercises such as Spending Reviews and the formulation of new strategies. They can contribute to the setting of policy and programme objectives, and can be used to demonstrate how those objectives are being met; and
- they can therefore provide accountability, by demonstrating how funding has been spent, what benefits were achieved, and assessing the return on resources. This can help to satisfy external scrutiny requirements and comply with sunset clauses and other formal requirements that make a link between evaluation and the continuation of the policy.

1.7 Good evaluation, and the reliable evidence it can generate, provides direct benefits in terms of policy performance and effectiveness, but is also fundamental to the principles of good government, supports democratic accountability and is key to achieving appropriate returns from taxpayers' resources. A good evaluation is therefore a normal and natural part of policy making and effective government and is a powerful tool available to the policy maker.

What factors affect how a policy should be evaluated?

1.8 Evaluations are a crucial (and in some instances mandatory – see Box 1.A) part of the policy cycle set out below and offer both strategic and practical benefits. Therefore, while it might be

tempting to do without an evaluation, or to 'muddle through' with a less formal, more subjective assessment of a policy's performance perhaps for time or resource related reasons, or the risk of a 'difficult' conclusion – such an approach is not without cost. A decision not to evaluate a policy, or only to evaluate it in a less formal or reliable way, is associated with a number of real risks:

- a policy which is ineffective might continue;
- overall adverse or costly impacts will be generated, now or in the future; or
- opportunities to improve the policy, or to save money or reinvest in other, more worthwhile projects might be missed.

1.9 Conversely, even if the policy is actually highly effective or generates good value for money, a substandard (or absent) evaluation will mean:

- Policy makers cannot justifiably claim that any positive outcomes they might observe were actually caused by the policy rather than by chance or were attributable to an alternative policy; and
- as a result, policy makers could not claim that their intervention delivered value for money, or had been demonstrated through sound analysis to be effective.

1.10 The key here is clearly the meaning of the phrase "good evaluation", what defines a good evaluation and what is necessary to achieve one. This is the subject of subsequent chapters of the Magenta Book. A wide range of factors needs to be taken into account when deciding what sort of evaluation is necessary and appropriate in any given case. These include:

- the nature of the policy, its objective scale, complexity, innovation, form of implementation and future direction;
- the objectives of the evaluation and the types of questions it would ideally answer;
- the timing of key policy decisions and the information on which they need to be based;
- the types of impacts which are expected, the timescales over which they might occur, and the availability of information and data relating to them and other aspects of the policy; and
- the time and resources available for the evaluation.

1.11 The choice of evaluation will often involve some trade-offs between these factors, which are considered further in Chapter 2. In some cases, it might be proposed that an intensive, rigorous evaluation is not justified, and a more limited, "lighter touch" evaluation is more appropriate. In others, it could be better to choose a more rigorous evaluation with a more restricted scope, since at least then the evidence obtained should be useful and reliable. However, such choices must be made in full recognition of the limits they are likely to place on what can subsequently be said on the basis of the results obtained.

1.12 The earlier that an evaluation can be planned in the policy development process, the more likely it is that it will be possible to consider these trade-offs and choose the most appropriate evaluation. The later in the policy process the evaluation is considered the fewer options there are for undertaking it. Judgement needs to be made during the development of the policy on the scale and form of evaluation that is required, which might even extend to considering whether policy implementation might be adjusted to make a stronger evaluation more feasible. This judgement will involve some technical issues and should therefore be made in consultation

with analytical specialists who can advise about the trade-offs involved and the implications of different choices.

Box 1.A: When is evaluation a formal requirement?

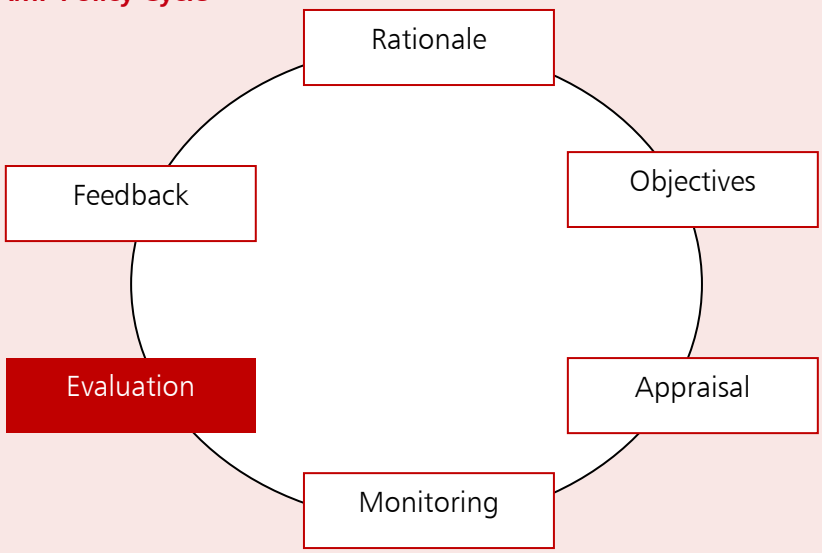
- There are a number of formal requirements to evaluate that need to be taken into account during the development of any evaluation, which might affect its scope, design and timing. Examples of when an evaluation might be a requirement include:
  - policies where a formal impact assessment was required and which are subject to Post-Implementation Review;
  - regulations containing a Sunset Clause or a Duty to Review clause; and
  - projects which are subject to a Gateway review also require a Post-Implementation Review as part of the Gateway 5: Benefits Realisation process.
- The National Audit Office (NAO) and the Public Accounts Committee (PAC) may examine the policy intervention being evaluated as part of their enquiries and would expect to see evidence that it was planned and implemented with due regard for value for money. Where the NAO undertakes a value for money study it will publish a report, which is likely to be the subject of a hearing of the PAC. The NAO's interest may include examining whether the intervention was subject to appropriate evaluation. (www.nao.org.uk)

How evaluation fits into the policy cycle

1.13 Evaluation is an integral part of a broad policy cycle that the Green Book formalises in the acronym ROAMEF. ROAMEF stands for Rationale, Objectives, Appraisal, Monitoring, Evaluation and Feedback. The ROAMEF cycle is presented in Chart 1.A. Though evaluation evidence can feed in throughout the whole policy cycle it is useful to highlight some of the key sections where evidence, including evaluation evidence can be used:

- appraisal occurs after the rationale and objectives of the policy have been formulated. The purpose is to identify the best way of delivering on the policy prior to implementation. It involves identifying a list of options which meet the stated objectives, and assessing these for the costs and benefits that they are likely to bring to UK society as a whole. The Green Book is the main source of guidance on appraisal;
- monitoring seeks to check progress against planned targets and can be defined as the formal reporting and evidencing that spend and outputs are successfully delivered and milestones met; and
- evaluation is the assessment of the policy effectiveness and efficiency during and after implementation. It seeks to measure outcomes and impacts in order to assess whether the anticipated benefits have been realised.

Chart 1.A: The ROAMF Policy Cycle



1.14 Chart 1.A suggests that these phases of the ROAMEF cycle occur in a stepwise fashion, but in practice this one-directional relationship rarely holds, the process is often iterative and there are significant interdependencies between the various elements. For example, data produced through monitoring activities are often used at the evaluation stage. In addition, evaluations can play a role in the policy development process – through, for instance, the use of pilots and trials – implying the presence of (potentially numerous) feedback loops at different stages of the cycle.

1.15 Therefore, whereas the simple ROAMEF policy cycle shows that an evaluation will take place after the policy has been implemented, evaluations can, in fact, occur at practically any other time. And importantly, decisions affecting and relating to any evaluation will almost always be taken much earlier in the policy process. Chapter 3 explains how what might seem minor aspects of the way a policy is formulated or implemented can have significant impacts upon the ability to evaluate it rigorously. It is important, therefore, to ensure that evaluation is considered and planned at the same time as the policy is being formulated so that these links can be recognised and accounted for.

2 Identifying the right evaluation for the policy

Key points

- Evaluations can be designed to answer a broad range of questions on topics such as how the policy was delivered, what difference it made, whether it could be improved and whether the benefits justified the costs.
- Broadly, these questions can be answered by three main types of evaluation. Process evaluations assess whether a policy is being implemented as intended and what, in practice, is felt to be working more or less well, and why. Impact evaluations attempt to provide an objective test of what changes have occurred, and the extent to which these can be attributed to the policy. Economic evaluations, in simple terms, compare the benefits of the policy with its costs.
- Understanding why an intervention operated in a certain way and had the effect it had generally involves combining the information and analytical approaches of the different types of evaluation and they should, therefore, be designed and planned at the same time.
- The choice of evaluation approach should be based on a statement of the policy's underlying theory or logic and stated objectives – how the policy was supposed to have its effect on its various target outcomes. The more complex the underlying logic, the more important it will be to account for other factors which might affect the outcome.
- Having a clear idea about the questions that need to be addressed and the required type(s) of evaluation at an early stage will help inform the design of the evaluation and the expertise required.

Introduction

2.1 This chapter discusses the different types of questions that evaluations can answer and provides a brief overview of the various types of evaluation that are possible. There are three broad classes of question which evaluation might be used to answer:

- How was the policy delivered?
- What difference did the policy make?
- Did the benefits of the policy justify the costs?

2.2 In most cases, there will also be considerable value in understanding why the policy was delivered in the ways it was, why the policy made the difference it did (or not), and how the costs and benefits were generated.

How was the policy delivered? Process evaluation

2.3 The question of how the policy was delivered is concerned with the processes associated with the policy, the activities involved in its implementation and the pathways by which the policy was delivered. These might vary quite considerably according to the nature of the policy in question, so there is no simple, generic characterisation of questions such as those that tend to be applicable in for impact evaluation.

2.4 However, using a practical example, such as the example of a policy of recruiting people onto a new training scheme to raise employment levels that is discussed at paragraph 2.7, questions might, for instance, seek to describe how individuals were recruited onto the scheme, what criteria were used to recruit them, and what the qualifications of training providers were. It might explore to what extent these factors varied across different parts of the country, and whether recruitment processes operated in favour of or to the detriment of particular groups, such as disabled people or those from particular ethnic groups. It could examine whether there were any difficulties or barriers to delivering the intervention as planned, and what steps were taken to increase course attendance. Box 2.A describes some of the approaches and methods which could be used to evaluate policy processes. Chapter 8 in Part B provides a more detailed description of process evaluation.

Box 2.A: How was the policy delivered? Process evaluation

Questions relating to how a policy was delivered cover the processes by which the policy was implemented, giving rise to the term "process evaluation". In general, process-related questions are intentionally descriptive, and as a result, process evaluations can employ a wide range of data collection and analysis techniques, covering multiple topics and participants, tailored to the processes specific to the policy in question.

Process evaluations will often include the collection of qualitative and quantitative data from different stakeholders, using, for example, group interview, one to one interviews and surveys. These might cover subjective issues (such as perceptions of how well a policy has operated) or objective aspects (perhaps the factual details of how a policy has operated). They might also be used to collect organisational information (for instance, how much time was spent on particular activities), although "administrative" sources (timesheets and personnel data, for instance) might be more reliable, if available.

Although essentially descriptive, these types of information can be vital to measuring the inputs of an intervention (which might not be limited to simple financial budgets, but might also include staff and other resources "levered in" from elsewhere) as well as the outcomes (surveys might be used to measure aspects of a scheme's participants' quality of life, for instance). This illustrates the practical link between process and impact evaluations, which often implies a need to consider the two together.

What difference did the policy make? Impact evaluation

2.5 Answering the question of what difference a policy has made involves a focus on the outcomes of the policy. Outcomes are those measurable achievements which either are themselves the objectives of the policy – or at least contribute to them – and the benefits they generate.

2.6 Questions under this heading might ask:

- What were the policy outcomes, were there any observed changes, and if so by how much of a change big was there from what was already in place, and how much could be said to have been caused by the policy as opposed to other factors?
- Did the policy achieve its stated objectives?
- How did any changes vary across different individuals, stakeholders, sections of society and so on, and how did they compare with what was anticipated?
- Did any outcomes occur which were not originally intended, and if so, what and how significant were they?

2.7 For example, a policy to recruit unemployed individuals onto a new training scheme which provides seminars to improve work skills might have the ultimate objective of reducing the costs of unemployment. It might attempt to do this by increasing the number of participants who receive and take up job offers, and increasing the duration of their employment. It might try and achieve this by improving participants' skills and qualifications, through seminar attendance and learning. Each of these measures – seminar attendance, number of job offers, duration of employment spells, the costs of unemployment, and so on – could be regarded as intended outcomes of the policy, and hence the subjects of the types of questions just described.

2.8 Questions relating to what difference the policy made concern the change in outcomes caused by the policy, or the policy "impact" – hence the term "impact evaluation", described briefly in Box 2.B. Issues around the reliability of impact evaluation results and how they are affected by the design of the policy are covered in Chapter 3, with further technical discussion provided in Chapter 9.

Box 2.B: What difference did the policy make? Impact evaluation

Impact evaluation attempts to provide a definite answer to the question of whether an intervention was effective in meeting its objectives. Impact can in principle be defined in terms of any of the outcomes affected by a policy (e.g. the number of job interviews or patients in treatment), but is most often focused on the outcomes which most closely match with the policy's ultimate objectives (e.g. employment rates or health status).

The key characteristic of a good impact evaluation is that it recognises that most outcomes are affected by a range of factors, not just the policy. To test the extent to which the policy was responsible for the change, it is necessary to estimate – usually on the basis of (often quite technical) statistical analysis of quantitative data – what would have happened in the absence of the policy. This is known as the counterfactual.

Establishing the counterfactual is not easy, since by definition it cannot be observed – it is what would have happened if the policy had not gone ahead. A strong evaluation is one which is successful in isolating the effect of the policy from all other potential influences, thereby producing a good estimate of the counterfactual. Sometimes the original business case for a policy might have made some estimates of this and forecast the difference the policy might make; this could be used in designing an evaluation. An evaluation might also be able to explain how different aspects of the policy contributed to the impact.

Whether a good impact evaluation is possible depends on features of the policy itself, the outcomes it is targeting, and how well the evaluation is designed. If a good evaluation is not possible, or the evaluation is poorly designed, the estimated counterfactual will be unreliable, and there will be uncertainty over whether the outcomes would have happened anyway, regardless of the policy. Then it will not be possible to say whether the policy was effective or not, and even if policy outcomes appear to move in desirable ways, any claims of policy effectiveness will be unfounded.

2.9 Clearly, there is overlap between the types of questions answered by process evaluation and those addressed through impact evaluation. Policy delivery can be described in terms of output quantities such as the numbers and characteristics of individuals that were recruited, how many training seminars were provided and how many individuals were in gainful employment after the training programme completed. But these are also measurable outcomes of the policy (although not necessarily outcomes which directly deliver benefits). This means that process evaluations often need to be designed with the objectives and data needs of impact evaluation in mind and vice versa. Using and planning the two types of evaluation together will, therefore, help to ensure that any such interdependencies are accounted for. The ability to obtain a convincing explanation will depend on the underlying "theory" of the intervention – that is, how the intervention was supposed to work (see section below on "What type of evaluation for the policy?")

Did the benefits justify the costs? Economic evaluation

2.10 A reliable impact evaluation might be able to demonstrate and quantify the outcomes generated by a policy, but will not on its own be able to show whether those outcomes justified that policy. Economic evaluation is able to consider such issues, including whether the costs of the policy have been outweighed by the benefits. There are different types of economic evaluation, including:

- cost-effectiveness analysis (CEA), which values the costs of implementing and delivering the policy, and relates this amount to the total quantity of outcome generated, to produce a "cost per unit of outcome" estimate (e.g. cost per additional individual placed in employment); and
- cost-benefit analysis (CBA), which goes further than CEA in placing a monetary value on the changes in outcomes as well (e.g. the value of placing an additional individual in employment). This means that CBA can examine the overall justification for a policy ("Do the benefits outweigh the costs?"), as well as compare policies which are associated with quite different types of outcome. CBAs quantify as many of the costs and benefits of a policy as possible, including wider social and environmental impacts (such as crime, air pollution, traffic accidents and so on) where feasible. The Magenta Book uses the very general term "value for money" to refer to the general class of CBA-based approaches, but it is important to recognise the more general scope of CBA which include those impacts which are not routinely measured in money terms. The Green Book provides more detailed guidance on CBA and the valuation of economic impacts.

2.11 Economic approaches value inputs and outcomes in quite particular ways, and it is crucial that the needs of any economic evaluation are considered at the design stage. Otherwise, it is very likely that the evaluations will generate information which, although maybe highly interesting and valid in itself, is not compatible with a cost-benefit framework, making it very difficult to undertake an economic evaluation.

Why did what happened occur?

2.12 Finally, there is the additional question of why what was observed about a policy's processes or outcomes occurred. In some limited cases, this might be of only secondary interest – so long as an intervention can be shown to work, the exact reasons why might be considered unimportant. In other cases, the particular evaluation technique adopted might not be capable of explaining the mechanisms involved. It is likely, however, that an understanding of why the policy generated the processes and outcomes it did will be desirable for a number of reasons, including:

- so that effectiveness and value for money can be improved by emphasising the most successful parts of the policy and minimising (and maybe stopping) those

which work less well. The understanding can also permit any factors which are hindering policy effectiveness to be addressed, including making the policy work better for those individuals or areas who benefited less than others, and avoiding any undesirable unintended consequences;

- so that policy scope and coverage can be successfully and effectively extended (e.g. through the national roll-out of a regional pilot). Future policy-making can be informed and improved through contribution to the evidence base around "what works"; and
- an understanding of the workings of a policy and the reasons for its success adds to the credibility of accountability and value for money statements, and improves transparency and decision-making, as outlined in Chapter 1.

What type of evaluation for the policy?

2.13 The preceding discussion has suggested a number of factors which should be considered when deciding what type of evaluation is appropriate for any given intervention. The first is the type of information required about the policy intervention, that is, the questions the evaluation needs to answer. Process and impact evaluations can sometimes consider similar issues and questions – a process outcome (e.g. the number of job interviews following a training scheme) can also be an "impact" outcome (e.g. the overall increase in the number of job interviews for the trainee group).

2.14 There is then the additional consideration of what sort of answers process and impact evaluations can provide. This chapter has portrayed the answers from process evaluations as more descriptive, and the answers from impact evaluations as more definite and in some sense "robust". This is because good impact evaluations attempt to control for all the other factors which could generate an observed outcome (that is, they attempt to estimate the counterfactual). But again, the distinction between the two is not as simple as this suggests. Chapter 3 provides more information about impact evaluations.

2.15 This is because the importance of controlling for these other factors depends on how many there are and how likely they are to affect the result of interest. If the relationship being examined between the policy and the desired outcome is a simple and direct one, there might be few intervening factors and the need to take account of them by estimating the counterfactual with some form of control group might be slight. In these cases, the more descriptive assessment provided by a process evaluation might be sufficient to give a robust answer about whether the policy delivered its desired outcome. However, if the relationship is complex, with many factors potentially affecting the outcome(s) of interest, a more descriptive approach is unlikely to be able to account for all these factors reliably, and a more formal attempt to estimate the counterfactual will be necessary.

How do evaluation questions relate to the underlying "logic" of the intervention?

2.16 Clearly, the complexity of the relationship(s) involved relates to the question being asked of the evaluation – and here the concept of the intervention "theory" or "logic model" is relevant. Logic models[1] describe the relationship between an intervention's inputs, activities, outputs, outcomes, and impacts defined in Table 2.A.

[1] For further information, the Department for Transport's Hints and Tips guide to logic mapping is a practical tool which can aid understanding and the process of developing logic models. Logic mapping: hints and tips, Tavistock Institute for Department for Transport, October 2010. http://www.dft.gov.uk/

Table 2.A: Definitions of the terms used in logic models[2]

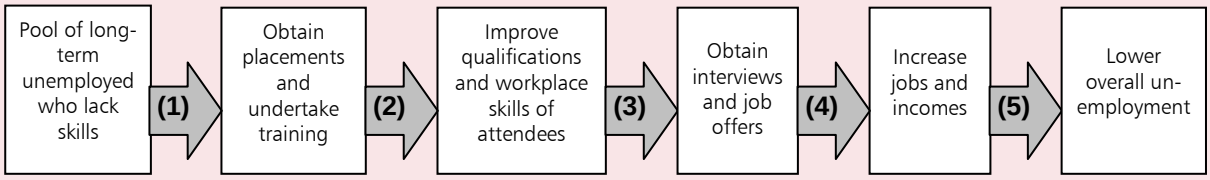
Term	Definition	Example
Inputs	Public sector resources required to achieve the policy objectives.	Resources used to deliver the policy.
Activities	What is delivered on behalf of the public sector to the recipient.	Provision of seminars, training events, consultations etc.
Outputs	What the recipient does 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.

2.17 Box 2.C presents a simplified logic model for a hypothetical intervention to reduce unemployment by increasing training. There are a number of steps in the intervention through which it is supposed to achieve its aims. As the number of steps increases, the complexity of the intervention also increases, as does the number of factors which could be driving any observed changes in outcomes, and the period of time over which they might be observed. But between any two given steps (e.g. link (1) in Box 2.C), the relationships are much simpler and there are fewer factors "at play". Hence, the importance of estimating a reliable counterfactual is reduced when the number of steps is lower, and increased as it rises. The relative suitability of process and impact evaluation for answering questions relating to how the intervention performed similarly is also likely to change with the number of steps.

[2] *Evaluation Guidance for Policy makers and Analysts: How to evaluate policy interventions*, Department for Business Innovation and Skills, 2011

Box 2.C: Formulating an evaluation: an example

As an example, suppose an evaluation is being planned for a job training scheme which is intended to provide placements for long-term unemployed people in companies where they can gain marketable skills and qualifications. The scheme aims to increase the number of interviews and job offers the participants receive, thereby increasing the number in jobs and their incomes. There might ultimately be a reduction in overall unemployment. A simplified intervention logic would be:



A number of evaluation questions arise from link (1) in the chain. For example, how were people recruited onto the scheme? What proportions were retained for the duration of their placement? For how long had they been unemployed before starting?

Link (2) might give rise to questions such as: what change was there in participants' skills and qualifications? Link (3) might describe the type and number of job offers obtained, and the characteristics of those participants obtaining them. But it might also involve assessing whether any improvement in skills contributed to participants gaining those interviews and job offers. Link (4) might measure the increase in the number and type of jobs, and the incomes of participants. There might also be interest in knowing whether the scheme generated genuinely new jobs, or whether participants were simply taking jobs that would otherwise have been offered to others.

Questions of interest under link (5) might include whether the scheme made any contribution to overall employment levels, either locally or nationally, taking account of economic conditions and trends. There might also be some attempt to measure the impact of the scheme on local economic performance and gross domestic product.

2.18 So using the example in Box 2.C, a process evaluation might be suitable for finding out which participants obtained which types of employment and what their characteristics were (link (4)). But this information would also be extremely valuable (and perhaps even necessary) to answer the question, "Did the training intervention increase participants' employment rates and incomes?", where the large number of possible factors affecting the result would mean that only impact evaluation is likely to be able to generate a reliable answer.

2.19 However, if the question is, "To what extent was the scheme successful in getting participants onto placements?", a process evaluation might be quite sufficient on its own. If participants were not accessing those placements previously, it might be reasonable to assume that any observed increase was down to the scheme. There might be some need to account for any "displacement" (e.g. participants switching from other placements they might have previously accessed), but if participants' training histories are reasonably known and stable, the chance that some other factor might have caused some sudden change in behaviour might be considered low. With such a simple question, although an impact evaluation might obtain a more robust answer, it might not add much more than could be achieved by a process evaluation.

2.20 Finally, the question might be, "What effect has the scheme had on overall unemployment?" (effectively links 1-5). The great many factors which determine overall unemployment (macro-economic conditions, the nature of local industries, and so on) would suggest that only an impact evaluation could feasibly secure an answer. However, with such a complex relationship, the chance of the effects of a single training scheme showing up in measures of even quite local employment could be very small, unless the scheme represents a very significant change of policy and injection of resources operating over a considerable length

of time. Even then, even a very strong, intensive impact evaluation might not be able to detect an effect amidst all the other drivers of the outcome.

Factors affecting the choice of evaluation approach

2.21 The choice of evaluation approach will therefore depend on a range of issues such as:

- how complex is the relationship between the intervention and the outcome(s) of interest. How important will it be to control for other drivers of the outcome of interest? If control is important, this might point more towards an impact evaluation approach. Simple relationships can often be investigated just as robustly by process evaluations;
- the "significance" of the potential outcomes in terms of their contribution to overall policy objectives. More limited, intermediate outcomes might be more readily evaluated robustly, but might not give a close or direct measure of the benefits of the policy;
- how significant the intervention is, in terms of the identifiable change in practice or increase in resources it represents. This will affect the extent to which the intervention could be expected to generate a large enough effect to "show up" amidst the other potential drivers. The distinction between projects, policies and programmes, strategy and "best practice" initiatives is relevant here, since these can vary significantly in terms of how much they represent distinct and identifiable interventions[3]; and
- how the intervention is implemented, and whether this facilitates or hinders the estimation of the counterfactual. This is discussed further in the next chapter.

[3] *Guidance for transport impact evaluations*, Department for Transport, March 2010, provides a fuller discussion http://www.dft.gov.uk/

3 Building impact evaluation into policy design

Key points

- Impact evaluations have special requirements which benefit from being considered during the policy design stage, because of the need to understand what would have occurred in the absence of the policy (generally through examining a comparison group of unaffected individuals or areas).
- Minor changes to policy design can dramatically improve evaluation options and quality. Conversely, failure to consider the evaluation early enough can limit those options and the reliability of the evidence obtained.
- When thinking about an impact evaluation technique such as randomised controlled trials and piloting should be considered. Where this is not feasible, alternative ways of implementing the policy, such as phased introduction and allocation by scoring, can strengthen evaluation significantly.
- These types of adjustments need not introduce delays or complications to policy implementation. However, if policy makers intend to by-pass these considerations due to other factors which are seen as over-riding, they should do so only after a full examination of the implementation options and the pros and cons entailed by each.

Introduction

3.1 This chapter looks in more detail at impact evaluations and at some of the minor changes that can be made in the policy design process to improve evaluation quality and reliability.

Thinking about impact evaluation when designing the policy

3.2 As discussed in Chapter 2, one of the keys to good impact evaluation is obtaining a reliable estimate of the counterfactual: what would have occurred in the absence of the policy. This is frequently a significantly challenging part of impact evaluation, because of the often very large number of factors, other than a policy itself, which drive the kinds of outcome measures relevant to public policy (e.g. increased employment, falling crime, reduced prevalence of obesity). There are various approaches to impact evaluation (sometimes termed research designs) which can be used to attempt to isolate the impact of the policy from all these other drivers. The success of these approaches largely depends on their ability to establish a counterfactual through obtaining what are called "comparison (or control) groups". This in turn is critically affected by the way the policy is "allocated", that is, who or where receives the policy and when.

3.3 In other words, the design and implementation of a policy affects how reliably it can be evaluated, and even quite minor adjustments to the way a policy is implemented can make the difference between being able to produce a reliable evaluation of impact and not being able to produce any meaningful evidence of impact at all. This chapter briefly explains the role of comparison groups in improving how well a policy can be evaluated, and then provides some

simple examples of how minor policy adjustments can improve the chances of a reliable evaluation. It finishes with a consideration of the factors which might be taken into account when deciding whether such adjustments might be appropriate

The role of comparison groups in identifying the impact of a policy

3.4 Research designs usually estimate the counterfactual by ensuring that there are some individuals, groups or geographical areas not exposed to the policy at some point during its implementation. A comparison can then be made between those who have been exposed to the policy and those who have not. A simple example of this is a medical drugs trial where one group of participants (the "treatment" group) receives a new drug and the other (the "comparison" or "control" group) receives a placebo. Who actually receives the drug or the placebo is decided by chance, through a formal randomisation process. Then, so long as the treatment and control groups are similar in all other relevant respects, they can act as comparisons for one another. If there is then any difference in observed outcomes between the two, it can reasonably be assumed (under certain technical assumptions) that the difference is due to the policy (treatment).

3.5 There are two obvious difficulties with applying this simple scenario to the public policy context. First, those areas or individuals who receive policy "treatment" in practice do tend to be different from those that do not in quite obvious and relevant ways. Crime reduction policies tend to be implemented more often and intensely in areas with higher crime rates. Individuals who enrol on employment assistance programmes tend to be those who have lower work skills, lower educational achievement and live in areas with poorer economic performance and prospects. Those who choose to stay in treatment for drug misuse tend to be those who are more motivated to improve their lives and reduce the costs of their drug problems. Then, the difference between the treatment and control groups will not just be that one received the intervention and one did not, but all of the other differences in underlying characteristics. The comparison will be between "apples and pears", and it will not be possible to tell whether differences in observed outcomes between the two groups are due to the intervention or something else.

3.6 Second, social policy interventions do not tend to be administered to the policy target group randomly, with no regard to perceived need, justification and so on. So there is not generally a group of untreated subjects who could have been eligible for the intervention but were purposely denied it. Those that do not receive an intervention tend to be those for whom it is deemed unsuitable, and will therefore be systematically different from those who are. So there is unlikely to be a readily available comparison group of non-treated individuals who are similar to those who do receive treatment.

What modifications might we make and why?

3.7 Controlling policy allocation – which individuals or areas receive which interventions, and when – can play a key role in successful impact evaluation by affecting whether there is a meaningful comparison group. Public policy interventions tend naturally to be allocated in ways which conflict with good impact evaluation, but there are some minor adjustments which can be made to policy allocation which can dramatically improve the feasibility of obtaining meaningful estimates of impact. A simple explanation of some of these adjustments is provided in Box 3.A.

3.8 At first glance, accommodating evaluation in these ways might appear to require compromising on policy effectiveness. There might be concerns that planning research designs will delay the launch of a policy. Not necessarily targeting those subjects in most "need" is sometimes claimed to be limiting the benefits recipients might gain. Holding back a comparison group of unaffected individuals is similarly sometimes claimed to be limiting the numbers able to

benefit. But there are strong counter-arguments against each of these points which should be recognised.

Box 3.A: What policy adjustments can improve evaluation chances? Some examples

Pilots

For interventions that are innovative, experimental or otherwise associated with a high degree of uncertainty, piloting is a recommended and often used way to introduce the policy. (A detailed review of pilots has been published by the Cabinet Office).[1] This allows the policy to be tried out and information collected before full-scale resources are committed. In terms of generating a comparison group, piloting works because not every potential subject is exposed to the policy immediately. However, there is still likely to be a temptation on the part of those owning or delivering the pilot to allocate the intervention to those deemed most in need or otherwise deserving of it, leading to the same 'apples and pears' problem as was described in paragraph 3.5. Piloting should therefore be combined with one of the other allocation mechanisms described below.

Randomisation and randomised control trials

How should the policy be allocated to pilot areas, or to individuals or institutions within those areas? The method offering the strongest measure of policy impact is randomisation, often in a form known as a randomised controlled trial (RCT). In an RCT, the allocation of individuals, groups or local areas to receive the intervention is determined by lottery or some other purely random mechanism. Carefully conducted, a RCT provides the clearest evidence of whether an intervention has had an effect. RCTs should therefore be near the top of the list of potential allocation mechanisms, especially for policies that are experimental in nature. However, it is often claimed that RCTs are not appropriate or possible for a variety of operational, underpinning logical or ethical reasons. Indeed, there are a range of factors which can make randomisation difficult to implement. For instance, it is not likely to be suitable for assessing the impact of changes in universal policies. (For example, it would not be feasible to change the law on the legal blood alcohol limit for a random selection of drivers).

Phased introduction and intermittent operation

A variant of randomised allocation is phased introduction, whereby all participants in the pilot receive the intervention, but sequentially over some period of time. The periods of time when some participants have received the intervention and others have not can then serve to generate a comparison group (though you still need to control in some way for other factors ongoing during the time delay). It is still preferable to use randomisation to determine the order in which participants receive the intervention, to avoid a situation where "the most deserving" or "most prepared" receive it first – this might be considered more acceptable within a pilot in which all participants are planned to receive the intervention eventually. Obviously, phased introduction need not be limited to pilots and can also be used for the roll-out of general (e.g. national) policies.

A further variant of the phased introduction approach might be termed intermittent operation, where interventions that are short term in nature are applied in bursts. This approach is only likely to be suitable for particular types of intervention which are appropriately flexible (advertising campaigns might be one example).

[1] *Trying it out – the role of "pilots" in policy-making*, Cabinet Office, 2003

Objective allocation rules

Where policies are targeted towards individuals, institutions or areas that have the greatest need (for example, prolific offenders, "failing" schools or deprived neighbourhoods), evaluation can be made much stronger (and the policy more transparent) by employing objective allocation rules (e.g. scoring systems or funding formulae) to determine who receives the policy. These policies can be evaluated effectively if these rules are well documented and applied. One approach is to assign a score to each offender, school, and so on, based on their level of need, so that those above a certain score then receive the policy, and those below do not. Comparison might then be made between subjects who received similar scores but who were just above and just below the threshold, or perhaps comparing those in just in scope of a policy with those just out of scope.[2] Waiting lists are an administrative approach to allocation which can combine the features of phased introduction and objective allocations rules (e.g. a scoring system to assess needs and hence treatment priority).

Measures of relative effectiveness

If a policy must be introduced everywhere simultaneously then it will not always be possible to obtain an estimate of the full policy impact. However, some modifications might allow an estimate to be made of the impact on effectiveness of changes in the level or intensity of policy exposure – that is, of one extent of implementation relative to another. In these cases, the level of exposure which a subject receives needs to be decided in a way similar to the approaches discussed here (e.g. randomly, or through a scoring system), to ensure that exposure is not tailored by the policy maker to match needs of the intervention target or participant

3.9 As regards the timing of policy launches, avoiding delays can simply be a question of sound project management – including preparing for the evaluation in parallel with the other activities necessary to set up the policy. Moreover, many of the allocation mechanisms described in Box 3.A could be said to represent rather minor modifications of practice which do not imply significant policy delays. Good impact evaluation can be compatible with quick policy timescales, so long as it is considered early enough in the development process.

3.10 In response to the claim that adjusting implementation will reduce effectiveness or that random allocation of the policy might raise ethical concerns that the policy would not be delivered to those most in need, at least with policies where there is a reasonable degree of uncertainty about outcomes or value for money, one of the principal reasons for undertaking an impact evaluation is to determine whether an intervention is effective or offers value for money at all. In these situations, it does not follow that temporarily restricting implementation or using random allocation will necessarily reduce policy effectiveness. It could just as easily be the case that overall effectiveness might actually increase, by avoiding resources being wasted subsequently on policies which do not work or do not offer good value for money.

3.11 Even when a policy is implemented initially in a restricted way (for instance, in the form of a pilot or phased introduction), it might still be targeted at those subjects deemed most in need, rather than through a less discretionary, more random process. This might be in an attempt to

[2] For example in the Department for Work and Pension's evaluation of the New Deal for Young People, those included in the policy scope (people aged 18-24) were compared with those out of scope (people aged 25 – 49) using a difference s in differences approach. See *Findings from the Macro evaluation of the New Deal for young people*, Department for Work and Pensions, 2002 http://www.dwp.gov.uk/

"appease" any persistent concerns about limiting effectiveness. However, if so, it should be recognised that there will be negative consequences for the eventual evaluation. Not only will it be made more difficult to achieve reliable results (for the "apples and pears" reason described in paragraph 3.5), but any results which are obtained will relate to the recipients of the restricted policy only, and will not be readily applicable to those areas or individuals which would come under a more widely rolled-out policy. This will make extrapolation more difficult.

3.12 It is clear that impact evaluation has certain special requirements. Often these can be met by taking some relatively simple steps during policy development. The risks discussed in paragraph 3.8 should be recognised, therefore, but not exaggerated or used as a routine excuse to avoid undertaking robust evaluation. Nevertheless, there might be occasions where there is pressure to implement a policy as quickly as possible, in a quite specific way, with little thought given to the implications for any subsequent evaluation. If this is the case, it is better for decisions to be made only once the implementation options have been identified and their implications for evaluation and evidence considered. In some cases, pressure to implement might simply reflect a lack of recognition of the negative consequences for the evaluation, or the ease with which evaluation needs can be accommodated.

4 What practical issues need to be taken into account when designing an evaluation

Key points

- Planning an evaluation involves identifying the evaluation audience and objectives, the appropriate evaluation type, the governance structure, the resources required and the timing. Developing an evaluation plan at an early stage will help to ensure that all the important steps have been considered.
- Any evaluation can require a variety of resource types, depending on the evaluation, including funding, staff management, procurement expertise, and analytical staff input.
- Evaluations need to be proportional to the risks, scale and profile of the policy, and this has implications for the type and level of resources required.

Introduction

4.1 Chapters 1 to 3 have introduced the key theoretical concepts of evaluation and what they mean for policy design. This chapter discusses some of the practical considerations when planning an evaluation, including when and how evaluations should or shouldn't be undertaken, and the resources required.

The main steps in the evaluation process

4.2 Planning and undertaking an evaluation will involve a number of steps and considerations. It can be helpful to develop a structured plan at an early stage, which ensures all aspects have been considered and helps guide the evaluation activity. This will normally be linked to the steps outlined in Table 4.A. Part B of the Magenta Book provides greater detail related to these steps.

Table 4.A: Steps involved in planning an evaluation

Steps involved in evaluation	Questions to consider
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?
Considering implications of policy design for evaluation feasibility	• 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?
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?
Identifying the data requirements	• At what point in time should the impact be measured? • What data are required? • What is already being collected / available? • What additional data needs to be collected? • Who will be responsible for data collection and what processes need to be set up?
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 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, and 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 a 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?

How to ensure an evaluation meets the requirements: governance and quality control

4.3 Quality control and quality assurance are crucial for any evaluation. Without these, the methods and results from the evaluation cannot be guaranteed to be of sufficiently high standard or fit for purpose. This means the resulting evidence is not robust enough to provide answers to the questions the evaluation was designed to resolve or to reliably inform the decision making process. Quality control can be described as follows:

- quality control ensures that the evaluation design, planning and delivery are properly conducted, conform to professional standards (such as ethical assurance), and that minimum analytical standards are adhered to;
- quality control will be informed by the governance community (e.g. a steering group), other stakeholders, the evaluation team, the manager of the evaluation within the commissioning body, external reviewers, and the commissioned research team where appropriate; and
- quality control will ensure consistency in data collection, methodology, reporting and interpretation of findings.

4.4 Without good quality control, the conclusions of an evaluation cannot be relied upon. Quality control and assurance should therefore be built into an evaluation. This will mean that any weaknesses in methodology, design, data collection and so on can be identified and understood early enough for changes to be made and adverse effects on results or reliability

avoided or reduced. This can be achieved by applying existing departmental quality criteria and processes for research and evaluation, and working closely with government analytical and evaluation specialists. The manager of the evaluation within the commissioning body should take responsibility for applying quality control criteria. The use of external assessors and/or peer review can also be useful and is often standard practice.

4.5 Four particular issues are often critical in managing an evaluation in a way that satisfies quality principles and criteria – ensuring independence, inclusivity, transparency and robustness:

- researcher independence and objectivity are essential for any evaluation. However, this does not automatically necessitate the use of external contractors or keeping the evaluation team at arm's length. This is because close interaction between the research team and policy colleagues while retaining independence and objectivity is important in delivering an effective evaluation;
- inclusion of recipients, delivery bodies or stakeholders – through a steering group, for example – enhances the potential learning from an evaluation and acceptance of its results, but it has to be actively managed as a continuous process of communication and engagement. This is likely to involve: improving awareness of the evaluation; obtaining feedback on research design; and communicating scoping, interim and final conclusions;
- transparency must be a feature of any evaluation but especially for a high-risk or innovative policy intervention. An evaluation plan can set out the evaluation objectives and questions, how the evaluation will be conducted, the timescale and how the findings will be acted upon. In turn, this will facilitate stakeholder engagement, allow the issues and risks to be identified and managed, and the delivery outputs and milestones to be agreed and documented. Evaluation reports should be published and contain sufficient technical detail for others to judge for themselves the robustness of the findings; and
- robustness in research plans and/or the final report is assessed against required analytical standards so that there is an assessment of a) whether the planned research is likely to provide robust evidence to answer the research questions and/or b) that the research findings and conclusions are presented and reported accurately and clearly.

Timing of the evaluation

4.6 Process evaluation is often able to identify when a novel policy is encountering initial difficulties in implementation, and so can be useful in ironing out these types of problems. This might mean that it is desirable for an impact evaluation to occur after a process evaluation, as analysts and policy makers can be more confident that the impact evaluation is measuring the policy itself, rather than the effects of delivery problems. However, this is likely to lead to a longer overall evaluation period. Some process and impact evaluations which follow a new policy as it develops can take years to complete, although useful results will usually be obtained throughout the study as well.

4.7 The timing of the evaluation will also be affected by the outcomes affected by the policy and of particular interest to the evaluation. Some impacts might take some considerable time (e.g. years) to appear, and it might be unfeasibly costly to incorporate these into an intensive process evaluation. An impact evaluation, undertaken some considerable time after the policy was implemented, might be the only feasible option for measuring these impacts, but might then be of less value in affecting the way the policy is implemented or rolled out.

4.8 Retrospective impact evaluations using existing data sources, will not generally suffer from the effects of implementation problems, and can sometimes be undertaken in a matter of weeks. However, the tendency to rely on administrative data will generally limit such an evaluation's ability to provide a rounded explanation of why and how any estimated impact actually occurred. Additionally, the timing of an evaluation might need to be aligned with specific requirements for review. Timetabling is particularly important where the evaluation is intended to inform a Sunset Review as it will need to be completed in time for any renewal or amendment legislation to be enacted (otherwise the legislation will automatically expire).[1]

What types of resources are likely to be needed?

4.9 Any evaluation will require significant input from both analysts and policy makers to ensure it is designed and delivered successfully. This is true for both externally-commissioned evaluations and those conducted in-house. A number of different types of resources will need to be considered and it is important to think early about these, ideally during the policy design process. The types of resources that are likely to be required are shown in Table 4.B.

Table 4.B: Types of resources employed in evaluation

Resource type	Description
Financial resources	A substantial part of the costs of an evaluation may be incurred after the policy has been implemented. Therefore, it is important to think about the financial resources required for the evaluation whilst planning the policy budget. Cost will be substantially lower if data can be used which already exist and/or are being collected through monitoring activities. Data collection exercises might need to be funded if the policy is novel or targeting unusual or hard-to-measure outcomes.
Management resources	Both internal and external evaluations will often require a dedicated project manager (with the specialist technical expertise to assure quality) who is responsible for: commissioning (for external evaluations); day-to-day management; advising the evaluation contractors and reacting to issues that develop. The level of input required will be greatest at key points (in particular, the design and commissioning stage), but this will be an ongoing resource requirement and should not be underestimated.
Analytical support	Due to the multi-disciplinary nature of many evaluations, it is important to consider the range of internal analytical specialists (such as social researchers, economists, statisticians, operational researchers, or occupational psychologists) who might need to be called upon for advice and to help design the evaluation approach and outputs. They can also advise on the effect of policy design on the feasibility of undertaking different types of evaluation. This can help ensure that the evaluation design will provide evidence to answer the research questions, and that, if necessary, appropriately skilled contractors are commissioned. Analytical input can also be useful in the steering of the project and in the quality assurance of outputs.
Delivery bodies	A successful evaluation will often depend crucially on the early and continued engagement and cooperation of the organisations and individuals involved in delivering the policy. It will be important to communicate what the evaluation seeks to address, what input will be required from them, and how they might benefit from the findings.

[1] Further guidance is provided in *Sunsetting Regulations: Guidance*, HM Government, 2011 http://www.bis.gov.uk

Wider stakeholders	The evaluation may also involve other stakeholders – for example, people and organisations directly or indirectly affected by the programme. The level of involvement and method of engagement will be specific to the policy and stakeholders in question, but may include inviting them onto a steering group, informing them about the evaluation, or including them as participants in the research.
Peer review	In order to ensure quality it may be necessary to have aspects of the evaluation peer reviewed. This is a requirement in some central government departments. Peer review might include the methodology, the research tools, and any outputs including interim and final reports.

What level of resource should be dedicated to the evaluation

4.10 Any evaluation needs to be proportionate to the risks, scale and profile of the policy. The feasibility and significance of obtaining robust evaluation findings will also be relevant and there may be certain circumstances where an evaluation is not feasible or appropriate, for example: when the specific policy can be regarded as part of a broader programme and evaluated at a higher level; when a policy is generally unpredictable or is changing; where costs for a full evaluation are prohibitively high; where there is a lack of consensus or clear direction about program goals; or where the evaluation findings won't be used.

4.11 It may also be argued, even for a relatively important intervention, that it is not possible to afford a full evaluation, in line with the recommendations in the Magenta Book. Certainly the guidance on proportionality should be taken seriously – evaluation research should only be carried out to answer questions in which there is genuine interest, and the answers to which are not already known.

4.12 But even after the overall affordability is queried, it is important to consider the opposite question – can one afford not to do a proper evaluation? Skimping on the research can have serious consequences. It is almost certain to be more cost-effective to conduct a robust evaluation, rather than have to repeat an evaluation because it was not adequately resourced. Furthermore, without a solid basis of evidence, there is a real risk of continuing with a programme which has negligible or even negative impact, or of not continuing with a cost-effective programme.

4.13 Judgement therefore needs to be made about the scale and type of evaluation that is required or possible and the trade-offs that this would require, including whether it should be commissioned externally or conducted (either partly or wholly) in-house. Table 4.C presents some of the factors to be considered when determining the level of resourcing required.

4.14 In some circumstances, a scoping or feasibility study may be conducted to support this decision making process. This can provide 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.

4.15 If it is still necessary to reduce evaluation budgets, the following additional questions may provide pointers to how this could be done without rendering the evaluation worthless:

- Is it possible to accept increased risk of drawing a false conclusion about the impact/cost-effectiveness of the intervention? Are all stakeholders content to accept the risk?
- Is it necessary to produce results for sub-groups of the targeted population? Or would the overall impact be sufficient? (The risk here is that a programme which works for some people but not all may be judged as ineffective)

- If face to face surveys are planned, could they be replaced with telephone interviews, postal or online surveys, possibly by reducing the amount of data collected?
- How long do outcomes need to be tracked for? Are there proxy or intermediate outcome measures that could be used? What are the risks of shortening the tracking period? (Very often, tracking over a longer period increases the costs.)

Table 4.C: Factors affecting appropriate resourcing of an evaluation

Factor	Explanation
Innovation and risk	High risk policies are likely to require robust evidence to understand both how they are working in practice and whether they are having the predicted impacts. In those cases where the innovative initiatives might offer "low cost solutions" evaluation resources might be "disproportionately" high but are still needed to demonstrate the scale of the returns on the policy investment.
Scale, value and profile	Large scale, high-profile, or innovative policies or policies that are expected to have high impact are likely to require thorough, robust evaluation to help build the evidence base on what works, meet accountability requirements, assess returns on investment and demonstrate that public money is well spent
Pilots	Pilot or demonstration projects, or policies where there is a prospect of repetition or wider roll out, require evaluation to inform future activities.
Generalisability	If it is likely that the findings will have a much wider relevance than the policy being evaluated, more resource may need to be allocated to ensure that the results can be generalised with confidence.
Influence	If the evaluation is capable of providing information which can have a large influence on future policy (for example, it can report at a strategic time-point and/or meet a key evidence gap) more resource is likely to be justified
Variability of impact	The effects of policies with highly uncertain outcomes or with significant behavioural effects are likely to be more difficult to isolate, and there is likely to be a greater case for conducting a more extensive evaluation.
Evidence base	Where the existing evidence base is poor or under-researched an evaluation is likely to require more resources in order to fill the gaps

Concluding remarks

4.16 Part A of the Magenta Book has given an overview of the key issues in policy evaluation, where it fits in the policy cycle, what benefits good evaluation can offer and some of the things to consider when planning and undertaking any evaluation activity.

4.17 Part B is aimed primarily at an analytical audience and therefore more technical, though it will be relevant too for interested policy makers. It covers in more detail some of the issues, challenges and steps to take in planning and undertaking an evaluation, including the setting of an evaluation framework, process and impact evaluation design and approaches to the interpretation and assimilation of evaluation evidence.