EVALUATING EU EXPENDITURE PROGRAMMES:

A GUIDE

Ex post and Intermediate Evaluation

XIX/02 - Budgetary overview and evaluation
Directorate-General XIX - Budgets
European Commission

First edition, January 1997

EVALUATING EU EXPENDITURE PROGRAMMES:

A GUIDE

Ex post and Intermediate Evaluation

First edition, January 1997

This guide was written by Nigel Nagarajan and Marc Vanheukelen of the Unit 'Budgetary overview and evaluation' in the Directorate-General for Budgets of the European Commission. The authors wish to acknowledge the helpful comments and suggestions received from colleagues in the various Commission services as well as from two independent experts.

Foreword

Evaluation is an essential part of modern public sector management practice. It is for this reason that the systematic evaluation of European Union expenditure programmes is one of the key components of the Commission's *Sound and Efficient Management 2000* initiative.

With the adoption of the Communication on Evaluation on 8 May 1996, the Commission outlined a series of concrete steps to promote best practice in this field. Whilst recognising that the operational services are first and foremost responsible for ensuring the evaluation of their own programmes, the Communication charged the financial services with the responsibility of developing a number of instruments of support. Among these instruments were instructional manuals for use by non-specialists, setting out the purpose, management and design of evaluation, some essential methodological questions and basic standards of good professional practice.

The present guide looks at intermediate and ex post evaluation of expenditure programmes. It is aimed at programme managers within the Commission services, as well as others who require a general introduction to the theory and practice of evaluation.

I hope that this guide will be both a useful contribution to the work of a wide range of services as well as a clear indication of the development of an *evaluation culture* within the Commission.

E. Liikanen

Member of the Commission
responsible for budgets
and administration

Table of contents

The first evaluation

In the beginning God created the heaven and the earth.

And God saw everything that He made. "Behold," God said, "it is very good." And the evening and the morning were the sixth day.

And on the seventh day God rested from all His work. His archangel came then unto Him asking, "God, how do you know that what you have created is 'very good'? What are your criteria? On what data do you base your judgement? Aren't you a little close to the situation to make a fair and unbiased evaluation?"

God thought about these questions all that day and His rest was greatly disturbed.

On the eight day God said, "Lucifer, go to hell."[1]

1. Introduction

Evaluation may be regarded by some as a diabolical exercise. However, if evaluations are well conducted, and if the results of evaluations are used by decision-makers, they can contribute to improved public programmes, as well as to increased transparency, accountability and cost-effectiveness.

Evaluation is not new. In some areas of EU activity, it has been established for several years now. Similarly, some Member States have a relatively long record of conducting evaluations and acting on their results. In other countries, both in Europe and in the rest of the world, it is increasingly being introduced.

The European Commission's *Sound and Efficient Management 2000* initiative (known as SEM 2000) includes the use of evaluation as a key element in improving the management culture of the Commission itself. **A key innovation of SEM 2000 is the requirement that systematic evaluation be introduced for all EU programmes**. This requirement was reinforced by the Commission in its Communication on Evaluation, which was adopted on 8 May 1996. Apart from setting out the obligations which services are required to meet in terms of evaluation, the Communication also provided for a number of instruments to be put at the disposal of services to assist them in this task. This present guide is one of these instruments.

This guide is intended to introduce officials to the main aspects involved in managing evaluations and to provide a broad overview of the main technical issues. It is aimed at the average programme manager within the Commission, rather than the evaluation specialist, i.e. someone who wishes to understand how to manage external evaluations or to perform basic internal evaluations of EU expenditure programmes. Evaluations of interventions without budgetary consequences are not covered, nor does the guide examine the evaluation of projects or policies. However, many of the concepts used in this guide will also be of interest to those who are concerned with evaluating projects or policies.

The main focus of this guide will be on *ex post evaluations* (conducted either on or after the completion of an intervention) and on *intermediate evaluations* (conducted during the implementation of an intervention). A separate guide will be issued on *ex ante evaluations* (conducted before the implementation of an intervention), which are sometimes referred to as *appraisals*.

The structure of the guide is as follows:

♦ Chapter 2 introduces some **key evaluation concepts and definitions**:

- what can be evaluated?
- what issues are raised by evaluations?

- who is involved in the evaluation?
- what types of evaluations are there?

♦ Chapter 3 is concerned with **preparing and managing evaluations**. It gives advice on:

- establishing a management structure for an evaluation
- preparing an evaluation project
- drawing up the terms of reference

♦ Chapter 4 deals with **conducting evaluations**. It introduces the reader to the main issues involved in:

- evaluation designs
- data collection techniques
- data analysis techniques

♦ Finally, chapter 5 covers **reporting and disseminating evaluations**. In particular, it looks at

- maximizing the use of evaluations
- the presentation of the evaluation report
- the dissemination of evaluations

The remainder of this first chapter addresses two main questions:

- what is evaluation?
- why are programme evaluations conducted?

It is followed by a general discussion of some of the special factors which should be taken into account in the evaluation of EU programmes.

1.1. What is evaluation?

1.1.1. Towards a definition of evaluation

What, then, is evaluation? This question is not as easy to answer as one might think. A number of different definitions of the term 'evaluation' have

been put forward, each with its own merits. Here is a selection of possible definitions:

> "*A critical and detached look at objectives and how they are being met*"[2]
>
> *"The examination of whether the legal, administrative and financial means put into place by a programme have enabled it to produce the effects it was supposed to produce and to attain the objectives which were assigned to it"*[3]
>
> "*A process which seeks to determine as systematically and objectively as possible the relevance, efficiency and effect of an activity in terms of its objectives*"[4]
>
> "*The systematic application of social research procedures for assessing the conceptualisation, design, implementation and utility of public programmes*"[5]
>
> "*An independent, objective examination of the background, objectives, results, activities and means deployed, with a view to drawing lessons that may be more widely applicable*"[6]
>
> "*The judgement of public interventions according to their results, impacts and the needs they aim to satisfy*"[7]
>
> *"The process of forming a judgement on the value of a programme"*[8]

Given that it is probably impossible to arrive at a single definition of 'evaluation' which will have universal appeal, we have chosen instead to identify some crucial elements which should normally characterise evaluations:

- **evaluations should be *analytical*** - they should be based on recognised research techniques;
- **evaluations should be *systematic*** - they require careful planning and consistent use of the chosen techniques;
- **evaluations should be *reliable*** - the findings of an evaluation should be reproducible by a different evaluator with access to the same data and using the same methods of data analysis;
- **evaluations should be *issue-oriented*** - evaluations should seek to address important issues relating to the programme, including its relevance, efficiency and effectiveness; and

- **evaluations should be *user-driven*** - this just means that successful evaluations should be designed and implemented in ways that provide useful information to decision-makers, given the political circumstances, programme constraints and available resources.

1.1.2. What evaluation is not

If it is not very easy to say precisely what evaluation *is*, it is easier to say what it *is not*.

Firstly, evaluations differ from ***scientific studies***. Both should be analytical, systematic and reliable. However, whereas scientists may undertake research in order to expand the sum of human knowledge and frequently confine themselves to one highly specialised discipline, evaluations are undertaken for more practical reasons. They are intended to be of *practical use* by informing decisions, clarifying options, reducing uncertainties and generally providing information about programmes within their own specific contexts. They also can draw on a wide range of analytical approaches.

Neither is evaluation the same as ***audit***. Audit is primarily concerned with verifying the legality and regularity of the implementation of resources (inputs) in a programme. Evaluation, on the other hand, is necessarily more analytical. It examines the programme from the point of view of society (defined from different possible perspectives). It looks at the validity of the strategy followed and whether objectives are appropriate given the problems to be solved and the benefits to be achieved. Auditors tend to have coercive powers, sometimes defined in legal texts, whereas evaluators must often rely on “good will” and the power of their arguments.

Audit has traditionally covered activities such as the verification of financial records (financial audit). A more recent innovation is known *as performance audit*, which is conceptually closer to evaluation. Performance audit is strongly concerned with questions of efficiency (of a programme’s direct outputs) and good management. Performance audit and evaluation share the same aim of improving the quality of programmes, but evaluation goes much further. It also looks at issues such as sustainability, relevance and the longer-term consequences of a programme.

Finally, evaluation must be distinguished from ***monitoring***. Monitoring examines the delivery of programme outputs (the goods and services produced by the programme) to intended beneficiaries. It is a continual process, carried out during the execution of the programme, with the intention of immediately correcting any deviation from operational objectives. Evaluation, on the other hand, is specifically conducted at a discrete point in the life cycle of a programme, and consists of an in-depth study. **Monitoring is of key importance to improving programme performance, and successful evaluation often hinges upon successful monitoring**, for example because monitoring often generates data which can be used in evaluation.

1.2. Why are programme evaluations conducted?

Programme evaluations are, of course, conducted with the general aim of improving programmes. They may also be conducted with the intention of identifying the effects of a programme on society, or to allow decision-makers to arrive at a judgement about the programme's value.

In this guide, we will move beyond these *general* reasons for conducting programme evaluations and distinguish between the following three *specific* reasons:

- to improve management;
- for reasons of accountability; and
- to assist in the allocation of budgetary resources.

Ex post and intermediate evaluations are often undertaken for **managerial reasons**, i.e. a concern with assessing and improving a programme's implementation. Typically, those involved in managing a programme need to know what its strengths and weaknesses are, how it can be improved, which aspects of the programme are functioning adequately and which aspects are not, and what are the reactions of clients, staff and others to the programme. This can also lead programme managers and decision-makers to reconceptualise the underlying problems which a programme is meant to address.

Accountability is another important reason, particularly in the EU context where there is increasingly a legal requirement for evaluation. Evaluation is of interest to both supporters and opponents of programmes, as well as to the ordinary citizen. Evaluations conducted for accountability purposes typically focus on the programme's impact (the degree to which it produces its desired outcomes) and its cost-effectiveness, and are meant to improve transparency.

Finally, evaluations can also be used to **improve the allocation of financial resources** within organisations. In the EU context, this reason is clearly linked to accountability. It has also assumed an increased importance in the light of the SEM 2000 initiative. Budgetary limitations together with a general concern with increasing value-for-money for the EU taxpayer encourage moves to transfer resources away from ineffective or irrelevant programmes towards programmes which are more efficient and more in tune with the evolving aims and objectives of the EU.

1.3. The evaluation of EU programmes

There are a number of important *special factors* which should be taken into account in the evaluation of EU programmes. We have summarised these as follows:

- **decentralised management** - the more decision-making is removed from day-to-day management and from the ultimate programme beneficiary, the greater the need at the centre for evaluation. In the case of many EU programmes, the distance (geographical and hierarchical) between decision-making, management and impact on the ground is rather large. Some programmes are administered by regional or local agencies in different countries. This creates a potential information gap. Evaluation can help to fill this gap.

- **subsidiarity** - article 3b of the Treaty on European Union (the Maastricht Treaty) states that

  > "In areas which do not fall within its exclusive competence, the Community shall take action, in accordance with the principle of subsidiarity, only if and in so far as the objectives of the proposed action cannot be sufficiently achieved by the Member States and can therefore, by reason of the scale or effects of the proposed action, be better achieved by the Community."

By shedding light on the value-added of different programmes, evaluation can contribute in a very meaningful way to the decision on whether it is appropriate for any given programme to be conducted at the EU level.

- **programme renewal** - in general, EU programmes tend to have a definite life span which is determined by the specific piece of legislation setting them up, i.e. the *legal base*. A new legal base is required if a programme is to renewed after this time. This allows for ineffective programmes to be discontinued and for effective programmes to be renewed or extended. Evaluation can be a useful input to this decision-making process.

Decision-making in the EU is complicated, and it inevitably has a strong political dimension. Evaluation cannot substitute for this process. Instead, it seeks to enlighten it.

The Commission has a key role to play in this process, and the intelligent use of evaluation will be an important element. Evaluations which are well planned and properly executed can be of great benefit to those with an interest in EU programmes. The Commission therefore has a responsibility to ensure that evaluations meet high professional standards, and that their results are properly reported.

With this in mind, this guide offers practical advice to programme managers who wish to benefit from evaluation with a view to improving and justifying their work.

Where to look for more information

A useful first source of information is the evaluation material produced by different services within the Commission. The unit or official responsible for evaluation within each Directorate-General or service should be able to advise on whether specific material is available for the evaluation of given programmes. The interested reader can then review some of the main evaluation texts, some of which are mentioned in the Select Bibliography at the end of this guide. These include Patton (1996), Rossi and Freeman (1993) and Mohr (1995). The distinction between evaluation, audit and monitoring is explained in MEANS (1995) and Conseil scientifique de l'évaluation (1996). A copy of the Communication on Evaluation adopted by the Commission on 8 May 1996 should be available from the unit or official responsible for evaluation within each Directorate-General or service.

2. Key concepts and definitions

In this chapter we will briefly examine some of the key concepts involved in evaluation. We will introduce these concepts by addressing the following important questions:

- what can be evaluated?
- what issues are raised by evaluations?
- who is involved in the evaluation?
- what types of evaluation are there?

The reader can also refer to Annexe 1 of this guide, which provides a glossary of technical terms.

2.1. What can be evaluated?

Evaluation is a very wide-ranging concept, and at a general level virtually anything can be evaluated. In practice, however, we usually find that the term is applied specifically to public sector interventions at one or more of the following levels:

- **project** - a single, non-divisible intervention with a fixed time schedule and dedicated budget.

  examples: a project to improve the irrigation system in a particular province of a developing country;
  a training course targeted at a specific group of unemployed workers in a particular region of a Member State.

- **programme** - a set of organised but often varied activities (a programme may encompass several different projects, measures and processes) directed towards the achievement of specific objectives. Programmes also tend to have a definite time schedule and budget.

  examples: the MEDIA programme designed to encourage development in the production, distribution and financing of television programmes;
  the LEADER Community Initiative (Structural Funds programme) designed to promote the development and structural adjustment of rural areas;
  the Phare programme designed to encourage economic transition in the associated countries of Central Europe and promote their eventual accession to the EU.

- **policy** - a set of activities, which may differ in type and may have different direct beneficiaries, which are directed towards common general

objectives or goals. Unlike projects and programmes, a policy is usually not delimited in terms of time schedule or budget.

examples: the Common Agricultural Policy;
the Common Foreign and Security Policy.

The present guide focuses on ***programme evaluation.*** There are particular aspects associated with the evaluation of projects and policies which are beyond the scope of this guide. However, many of the points raised in the discussion of programmes will also be of interest to those who are concerned with evaluating projects or policies. The guide will also be relevant to those who are interested in so-called *thematic evaluations*, i.e. evaluations of one or more themes common to several different programmes or activities (e.g. effects on the environment or on small and medium-sized enterprises).

2.2. What issues are raised by evaluations?

2.2.1. The programme and its intervention logic

The evaluator must describe the programme being evaluated. This includes determining the *needs* which it seeks to address, the *objectives* which have been set and the *indicators* which allow us to judge its performance. However, evaluations must go beyond merely describing the programme. **A key task for the evaluator is to examine the validity of the programme's *intervention logic***. We will briefly discuss each of these concepts.

Programmes are always conceived with a given set of ***needs*** in mind. These needs are *the socio-economic problems which the programme seeks to address, expressed from the point of view of its particular target population(s), i.e. its intended beneficiaries*. Consider a programme aimed at reducing unemployment among the long-term unemployed (the target population). This group may suffer from a lack of relevant job skills (the socio-economic problem that has to be addressed). Hence, there is a need to improve their employment opportunities.

In order to tackle the socio-economic problems and address the needs of the target population, programmes pursue certain ***objectives*** (*desired effects*). For expenditure programmes, objectives can be expressed either in terms of:

- ***outputs*** (the goods and services funded and directly produced by the programme);

- ***impacts*** (the socio-economic changes brought about by the programme).

To emphasise this distinction, we can say that

outputs *are the things the programme produces*,
impacts *are the effects the programme induces*.

We will further divide ***impacts*** into:

- ***results*** (the initial impact of the programme); and
- ***outcomes*** (the longer-term impact of the programme).

Corresponding to the distinction between *outputs*, *results* and *outcomes*, there are three types of objective:

- ***operational objectives*** - are expressed in terms of **outputs** (e.g. to provide professional training courses to the long-term unemployed);
- ***specific objectives*** - are expressed in terms of **results** (e.g. to improve the employability of the long-term unemployed by raising their skill level). NB a programme may have different target populations corresponding to its different specific objectives; and
- ***general objectives*** - are expressed in terms of **outcomes** (e.g. to reduce unemployment among the previously long-term unemployed).

How do we know if a programme has met its various objectives? In order to judge the performance of a programme in this respect, we need to rely on indicators. For our purposes, an ***indicator*** is a *characteristic or attribute which can be measured to assess a programme in terms of outputs or impacts*. By necessity, indicators are simplifications of a more complex reality. They can be either quantitative (e.g. per capita GDP) or qualitative (e.g. trainees' opinions of the usefulness and relevance of a training course).

Output indicators are normally straightforward, since the programme managers will usually have information on the goods and services which the programme produces. This is, after all, the task of the monitoring system. Impact indicators may be more difficult to derive, e.g. because it may not be easy to determine what effects have genuinely been caused by a programme or because it would be costly and time-consuming to measure these effects *directly*.

For these reasons, it is often appropriate to rely on indirect indicators. Consider the example of a programme designed to raise literacy levels across an entire country. It may be costly or time-consuming to assess the reading skills of the population across different points in time. Instead, one could rely on figures for the sales of newspapers and books, bearing in mind that there can also be problems in interpreting indirect indicators. For example, the sales of newspapers and books will also be affected by competition from radio and television.

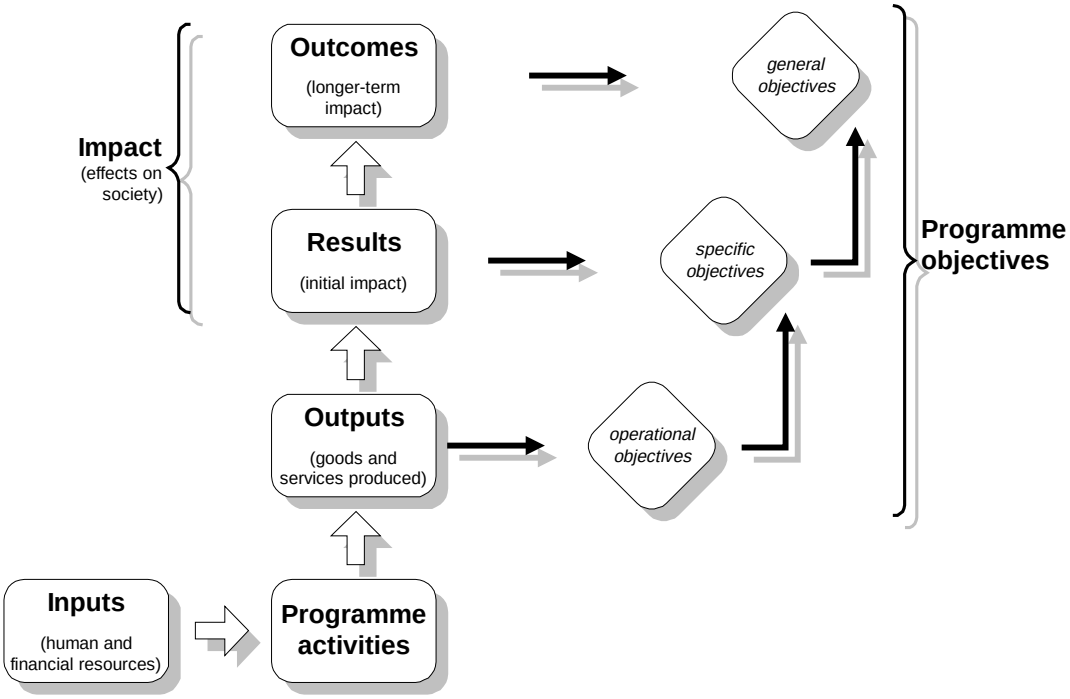
This leads us to the concept of the programme's ***intervention logic***. This refers *to the conceptual link from a programme's inputs (the human and financial resources devoted to it) to its outputs and, subsequently, to the achievement of the programme's results and outcomes*. A comprehensive

evaluation will systematically examine the validity of this logic. Figure 2.1. below shows how one should conceptualise the intervention logic of a typical programme (NB the intervention logic of a project or a policy will be somewhat different).

In plain language, the intervention logic of a programme is simply ***an explanation of what the programme is supposed to achieve and how it is supposed to achieve it***.

The examination of the programme's intervention logic will be of central importance in most evaluations. **The evaluator needs to ask *how* the inputs devoted to the programme lead to the various outputs, and *how* these outputs, in turn, lead to the results and outcomes which are expected of the programme**. In other words, how does the programme achieve its specific objectives, and how do the specific objectives contribute to the attainment of the general objectives?

Figure 2.1. The intervention logic of a programme



Typically, a programme's intervention logic will contain hidden assumptions (about causal links between the programme and its supposed effects and about how the programme influences, and is influenced by, other factors). *An important task is to identify these hidden assumptions so that they can be critically assessed by the evaluator*.

2.2.2. Key evaluation issues

Once the evaluator has described the programme and examined its intervention logic, he will typically move on to address several, if not all, of the following key evaluation issues:

- ***relevance*** - to what extent are the programme's *objectives* pertinent in relation to the evolving *needs* and priorities at both national and EU level.

- ***efficiency*** - how economically have the various *inputs* been converted into *outputs* and *results*?

- ***effectiveness*** - how far have the programme's *impacts* contributed to achieving its specific and general objectives?

- ***utility*** - how do the programme's *impacts* compare with the *needs* of the target population(s)?

- ***sustainability*** - to what extent can the positive changes be expected to last after the programme has been terminated?

Figure 2.2. below shows how each of these key evaluation issues relates to the programme being evaluated[9]. The diagram is divided into three different levels. The lowest level is that of **judgement**. Each of the above five issues are the responsibility of the evaluator. He has to use sound analytical techniques to arrive at judgements as to each of them.

The second level is that of the **programme** itself. The *objectives* behind the programme are what motivates it. To meet these objectives, human and financial resources (*inputs*) are devoted to the programme, and are allocated to various programme *activities*. This process leads to the generation of goods and services by the programme, which are its *outputs*.

The highest level is that of **socio-economic problems**. It is at this level that we should consider the *needs* of the target population and the particular *problems* which the programme is designed to address. The programme's *results* and *outcomes* are placed at this level because they affect these needs and problems. The dashed lines serve to indicate that the three levels are conceptually distinct from one another. For example, it may be difficult to identify what effects are genuinely caused by a programme and to separate these effects from the myriad of other influences on the socio-economic problems.

Let us now return to the level of judgement, and examine each of the *key evaluation issues* discussed above. The importance of the ***relevance*** criterion is that it can lead to decisions about whether a programme should be allowed to continue in its current state, should be altered significantly, or merely allowed to lapse without being renewed. When examining the relevance criterion, the evaluator will typically be asking whether broad changes in society have altered the rationale for a programme, or may do so in the

future. The discussion of *future relevance* normally entails an examination of alternatives to the programme.

As we have seen, ***efficiency*** compares inputs (resources) with the programme's outputs (the goods and services it provides) and results (its initial impact). An examination of efficiency involves asking: could the same benefits have been produced using fewer inputs? Alternatively, could the same inputs have produced greater benefits? Discussions of efficiency necessarily entail comparisons with alternatives to the programme. The main difficulty in this area is therefore the choice of appropriate benchmarks. The evaluator will need to specify which benchmarks the efficiency of a programme is being measured against. A difficulty can arise when there are no comparable programmes and the evaluator has no previous experience with similar programmes. Chapter 3 includes a more in-depth discussion of benchmarks.

Another important point to bear in mind is that even if a programme is efficient, it can still be poorly designed. This brings us to the discussion of ***effectiveness*** (comparing a programme's impacts with its objectives). It is worth remembering that in the case of such poorly designed programmes, objectives may not have been stated sufficiently clearly or may even be missing altogether. The evaluator may therefore be called upon to transform vague or general goals into verifiable objectives.

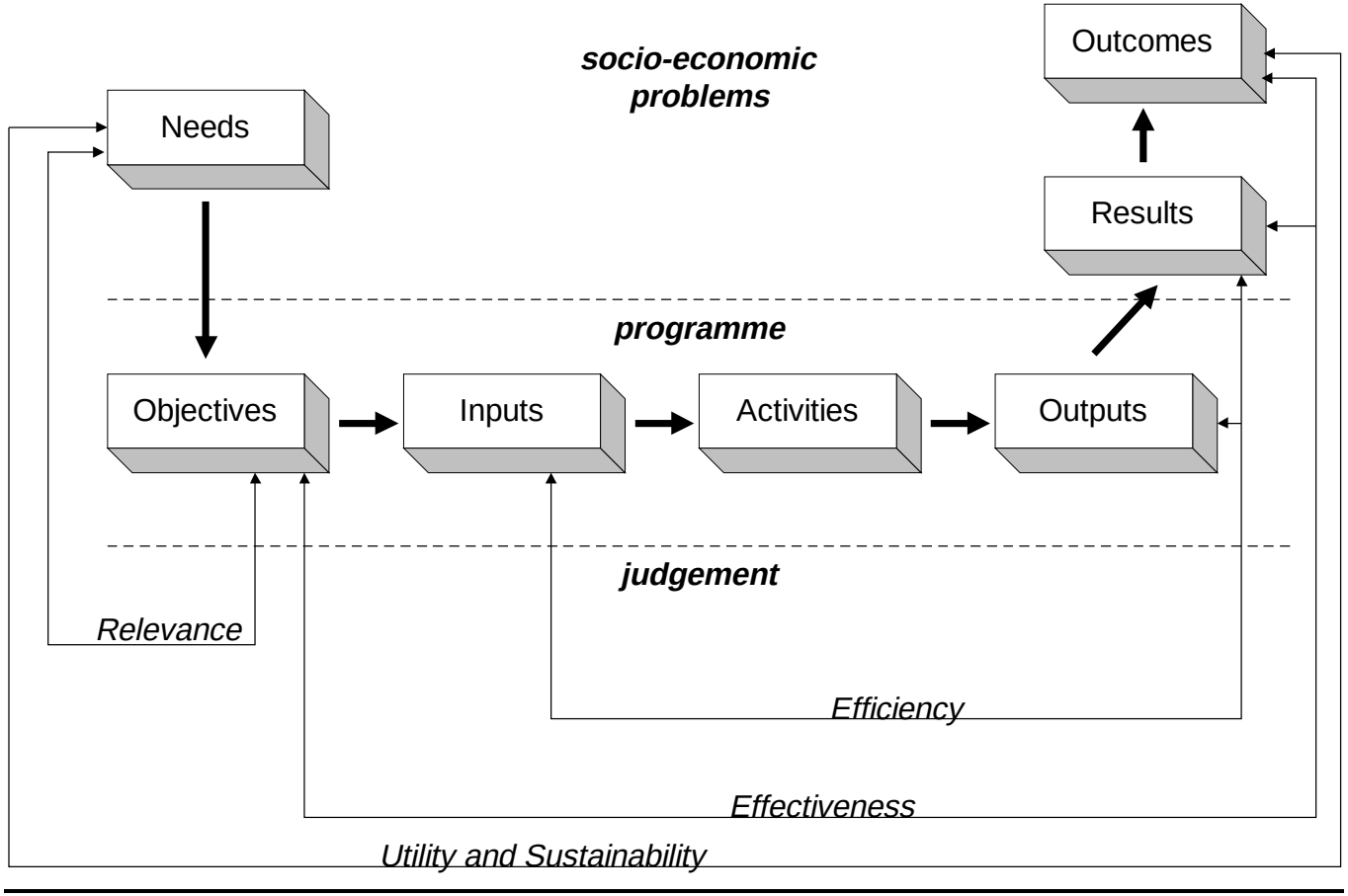
In addition, it must also be remembered that *effectiveness is concerned with only one aspect of a programme's impact*: the positive, expected effects. A programme may also have positive, unforeseen effects as well as negative effects (both expected and unforeseen). An evaluator will typically want to go beyond effectiveness in order to assess a programme's total impact, even if establishing causality is often difficult.

In order to assess the total impact of a programme, an evaluator is required to infer causality between it and the positive and negative, expected and unforeseen changes which have come about. Attributing causality is a key problem in the design of evaluations. **Other possible explanations for the effects which are to be attributed to the programme must be identified and, wherever possible, eliminated so that the evaluator can show that the positive effects would not have arisen anyway**. Causality is examined in more detail in Chapter 4.

This brings us to the concept of ***utility***, in which we compare the impact of a programme with the needs which gave rise to it. **Programmes will only be useful if they manage to bring about changes in society which are beneficial given the needs of the target population**.

When examining the utility of EU programmes, it is legitimate to ask whether they are consistent with the principle of *subsidiarity*. Is the programme useful compared to existing programmes at regional or national level? Would the programme be more useful if run at a different level of administration?

Figure 2.1. Key evaluation issues



A particular problem with the concept of utility is that, given that there is a multiplicity of different interests associated with public interventions, it is difficult to arrive at a universally acceptable definition of needs. Unemployed workers may define their own retraining needs in quite a different way from those administering the retraining programme.

Finally, we come to ***sustainability***, which is closely related to utility. Even if a programme generates benefits which are in tune with the needs of its target population, it may be of little value unless these benefits are still being enjoyed at some stage in the future. Sustainability is therefore concerned with what happens after a programme has been completed. For example, there is little value in training unemployed workers in skills which are likely to become obsolete after a few years. *If a programme is to be of lasting value, it must generate sustainable benefits for its target population(s)*.

To return to figure 2.2. above, it will be seen that each of the issues which we have examined in this section involves the evaluator making a technical evaluation **judgement** about either:

- the **programme** itself - *relevance* and *efficiency* (when it simply looks at how inputs are converted into outputs); or,

- the **programme** and the **socio-economic problems** it seeks to address - *efficiency* (when it compares inputs with results rather than merely outputs), *effectiveness*, *utility* and *sustainability*.

We have drawn a conceptual distinction between the level of judgements, the level of the programme's operation and the level of socio-economic problems. This distinction is very important. ***A programme's outputs should be directly identifiable, but identifying its results and outcomes may be far more difficult***. Both results and outcomes arise through a series of potentially complex interactions between the programme and society. Furthermore, there are also likely to be a myriad of other factors at work. The evaluator needs to rely on sound analytical techniques to establish how the programme changes society.

2.2.3. Other important issues

Moving beyond the five key evaluation issues which were discussed above, an evaluation may also be concerned with addressing other important issues. These other issues will be largely a function of the particular features of the programme (or programmes) being evaluated. Thematic evaluations, for example, focus on one or more common themes in the evaluation of different programmes or activities (e.g. effects on the environment or on small and medium-sized enterprises).

Three issues which are particularly relevant to evaluations of public expenditure programmes are *deadweight*, *displacement* and *substitution*. We will briefly discuss each of these.

Deadweight *is defined as effects which would have arisen even if the public expenditure programme had not taken place*. Deadweight usually arises as a result of inadequate programme delivery mechanisms (the organisational

arrangements which provide the goods and services funded by the programme to its intended beneficiaries). In particular, these mechanisms fail to target the programme's intended beneficiaries sufficiently well. As a result, other individuals and groups who are not included in the target population end up as recipients of benefits produced by the programme. For example, a retraining programme aimed at the long-term unemployed may benefit some people who would have undertaken retraining even without the programme (e.g. by pursuing higher education or privately-financed training programmes) and may not be genuinely long-term unemployed.

For many programmes, deadweight may be to some extent inescapable. However, it is an important issue in evaluating expenditure programmes because there is a concern both with identifying the extent to which the programme is meeting the needs of its target population and with reducing waste and inefficiency in public expenditure. *It should be clear that the issue of deadweight is closely related to that of efficiency, discussed above:* ***deadweight is really a special case of programme inefficiency***.

Displacement and substitution are two closely related terms which are used to describe situations where the effects of a programme on a particular individual, group or area are only realised at the extent of other individuals, groups or areas. Consider, for example, the case of a programme to provide employment subsidies. In a firm which benefits from this programme, subsidised workers may take the place of unsubsidised workers who would otherwise have been employed by that firm. This is known as ***substitution***. Alternatively, a firm benefiting from the employment subsidies may win business from other firms which do not participate in the scheme. Thus, the jobs created in the participating firm may be partly or wholly offset by job losses in other firms. This is known as ***displacement***.

Displacement and substitution are really special cases of negative programme effects, mentioned above. An analysis of the total impact of a programme needs to take these negative effects into account.

2.3. Who is involved in the evaluation?

The evaluator, the person directly responsible for conducting the evaluation, needs to be aware that there are whole range of different individuals and groups who may have a legitimate interest in his work. The term ***stakeholders*** is sometimes used to describe *the various individuals and organisations who are directly and indirectly affected by the implementation and results of a given programme, and who are likely to have an interest in its evaluation*.

A list of stakeholder groups who may either directly participate or have an interest in the evaluation process could include the following:

- policy-makers and decision-makers;
- those responsible for the evaluation of the programme;
- the target population of a programme;

- programme managers and administrators; and
- other individuals and groups with a legitimate interest in the programme.

The evaluator will normally be chosen by, and be directly responsible to, the *evaluation sponsors*. They have overall responsibility for the evaluation. In the case of EU programmes, this will normally be the managing Directorate-General or service within the European Commission. Chapter 3 provides a more detailed discussion of the relationship between the evaluator and the various stakeholder groups.

Those who write evaluation reports must demonstrate an understanding of the different information needs of the various stakeholder groups, as well as the relative importance of the different stakeholders at various stages of the evaluation. This is discussed in more detail in Chapter 5.

2.4. What types of evaluations are there?

This section is divided into three parts. Firstly, the distinction between formative and summative evaluations is explained. Before considering whether to conduct an evaluation, it is important to be clear about whether it will be mainly formative or mainly summative. Secondly, we will examine the distinction between intermediate and ex post evaluations. Thirdly, there is a discussion of the differences between internal and external evaluations.

2.4.1. Formative and summative evaluations

The type of questions which an evaluation asks will to a large extent be determined by who its intended users are and what their reasons are for requiring it. To illustrate this, it is helpful to distinguish between:

- **formative evaluations** - these are concerned with examining ways of *improving and enhancing the management and implementation of programmes*. Formative evaluations tend to be conducted for the benefit of those managing the programme with the intention of improving their work; and

- **summative evaluations** - these are concerned with determining the essential effectiveness of programmes. Summative evaluations tend to be conducted for the benefit of external actors (groups who are not directly involved in the management of a programme), for reasons of *accountability* or to *assist in the allocation of budgetary resources*.

Although the distinction between formative and summative evaluations may appear to be clear-cut, in practice it is often blurred. A general concern with improving public programmes usually requires a combination of both approaches. In the present guide, we will be mainly concerned with summative evaluations, or at least evaluations with a strong summative focus.

2.4.2. Intermediate and ex post evaluations

The present guide focuses on intermediate and ex post evaluations. The difference between the two is mainly a question of timing.

- **intermediate evaluations** - are conducted during the implementation of a programme.

- **ex post evaluations** - are conducted either on or after the completion of an intervention.

In many cases, intermediate evaluations often focus on a programme's outputs and do not attempt a systematic analysis of its impact. They therefore tend to rely quite strongly on information provided by the monitoring system. Intermediate evaluations may also tend to have a formative bias, e.g. a concern with improving the programme's delivery mechanisms. In other cases, intermediate evaluations do look at impact, but only in a limited way.

Ex post evaluations are more likely to be summative in nature, and are often conducted with the express intention of analysing a programme's impact. However, since the information needed to assess a programme's impact may often not be fully available until several years after the end of the programme, even ex post evaluations can be limited in the extent to which they can provide a complete assessment of impact. Since many EU programmes are replaced by successor programmes of a different generation, there can also be a legitimate interest in formative issues at the ex post stage.

2.4.3. Internal and external evaluations

A key decision in any evaluation plan is whether to opt for an *internal evaluation* or an *external evaluation*. These two terms can be defined as follows:

- **internal evaluations** - are performed by members of the organisation which is conducting the activity being evaluated

- **external evaluations** - are performed by persons outside the organisation managing the intervention.

In the EU, by far the greater part of evaluation undertakings are contracted out to external consultants, and this is especially the case for ex post and intermediate evaluations. There are, of course, tremendous advantages in using external consultants. They should normally be able to express an independent viewpoint on EU programmes. In other words, external evaluators should be able to evaluate *objectively*. External evaluators also tend to have an expertise in evaluation practice, and contracting out the evaluation task to an external consultant may be the most practical and cost-effective solution.

Internal evaluations can also have their benefits. In particular, internal evaluations may promote 'learning by doing' since managing services themselves are closely involved in questioning the 'how' and the 'why' of their activities. However, for many intermediate and ex post evaluations, internal evaluations may not be practical, cost-effective or even desirable. For example, it may be difficult to convince other stakeholders that an internal evaluation has

been conducted objectively. There is therefore a reliance on external evaluations in many Commission services.

In order to ensure that external evaluations are conducted properly, services must pay particular attention to drafting the terms of reference. Furthermore, unless there is proper supervision of the external evaluator during the conduct of the evaluation by the evaluation sponsors, a number of problems can arise. For example:

- evaluation reports prepared by external consultants may produce **misguided recommendations**, because the report has been prepared by people with insufficient knowledge of the EU organisational or political context; and

- there may be **problems of communication**: external evaluators may be too far removed from the chain of management for their findings to be taken into account;

On the other hand, ***there is an important need to ensure that the supervision of external evaluators by the sponsoring service does not compromise the evaluator's independence***. A steering group can be of tremendous help in this respect.

There are obviously trade-offs to be considered when choosing between an internal and an external evaluation. The technical competence and supposed independence of an external consultant should be weighed against any potential advantages from conducting an evaluation in-house. Chapter 3 includes more practical advice on the choice of the evaluator.

Where to look for more information

Annexe 1 of this guide contains a glossary of technical terms.

---

3. Preparing and managing evaluations

Evaluation is sometimes called “applied common sense”. Unlike common sense, however, evaluations have to be well prepared and properly managed.

- when evaluations are not well prepared, there is a danger that they can be carried out inefficiently. It is very easy to ignore important questions (is the programme at all evaluable? what is and what is not to be evaluated? for what purpose? how? by whom? for when? with what resources?) before evaluations are launched. These questions may seem obvious after the evaluation has taken place, but they need to be properly addressed beforehand.

- when evaluations are not well managed, there is a similar danger. Even in an evaluation which is well prepared, things can go wrong or circumstances can change in an unforeseen way. Sound management practices have to be followed when this happens.

Evaluations which are not well prepared and well managed may also suffer from *problems of credibility*. This reduces the chances of obtaining a broad endorsement of conclusions and recommendations from the interested parties (the stakeholders). In these circumstances, *the evaluation will only be of limited use*.

In this chapter, we will discuss each of the main components involved in preparing and managing evaluations. These are shown in Box 3.1. below.

Box 3.1. The main components in preparing and managing evaluations

- *establishing a* ***management structure*** - this involves setting up a clear hierarchy, which allows for overall management of an evaluation.

- *elaborating an* ***evaluation project*** - this involves a sequence of logical steps from the basic problems and interests motivating the evaluation to the questions which can be addressed in an analytically acceptable way.

- *drawing up the* ***terms of reference*** - this involves defining the relationship between those responsible for commissioning the evaluation (the evaluation sponsors) and those responsible for actually conducting it.

We will examine each of these components in more detail.

3.1. Establishing a management structure

A ***management structure*** *allows for overall management of an evaluation, and, in particular, the evaluation project*. An efficient management structure should ensure that the evaluation report is of high quality, available in good time and produced at a justifiable cost. The chief task of the management structure is to elaborate the evaluation project (see section 3.2. below) and define the terms of

reference for the evaluation (see section 3.3 below), in particular if the latter is entrusted to an external expert.

As a minimum, such a management structure must involve:

- the programme management; and
- the unit, sector, or official inside the same DG or service responsible for evaluation.

However, *it is often helpful to widen the management structure by creating a* ***steering group***. This applies especially when programmes are of major budgetary significance, or of a controversial nature, or when the evaluation's focus is not simply confined to the implementation of the programme but also looks at the programme's effectiveness and future relevance.

Apart from the DG or service in charge of the programme, such a steering group usually includes other DGs and services, such as those with a specific interest in the programme or with a general evaluation responsibility. There might also be representatives of the Council of Ministers and European Parliament in their capacity as branches of the legislative and budgetary authority. Significant stakeholders outside the EU institutions may also be represented. In addition, there could be independent experts, with the task of assisting in the elaboration of complex evaluation projects or ensuring a degree of quality control on the evaluation itself.

A key question to be addressed when setting up a steering group is whether or not to include representatives of those who are responsible for the actual implementation of the programme (e.g. implementing agencies). If they are included, it is important to ensure that the independence of the evaluation is not compromised.

A steering group has several advantages:

- it encourages active involvement in the evaluation by the various stakeholders;

- it reduces the chances that programme managers will become too closely associated with the evaluator, thus compromising his independence; and

- it allows for quality control of the evaluation by experts.

Creating a steering group helps to ensure that the evaluation is viewed as an *inclusive process*. Stakeholders are then more likely to have confidence in the evaluation's conclusions and recommendations, especially if they have had the opportunity to influence the design of the evaluation. However, *it is important to ensure that the steering group does not become too large*. It may then lose its role as a management body and degenerate instead into a negotiation forum, threatening the impartiality of the exercise. Evaluation must never become entangled with negotiation.

Regardless of whether or not a steering group has been created, it is the responsibility of the management structure to deal with problems or changes in circumstances which can arise once an evaluation is underway. These can include:

- disagreements between the steering group and the evaluator on some basic aspect of the evaluation design. It is not uncommon for steering groups to ask for the impossible, e.g. evaluations which are both formative and summative and which will pronounce upon the effectiveness of a programme despite the fact that the data necessary to form such judgements will not be available for several years. Ideally, these sorts of problem can be avoided at the outset if the evaluation project is properly elaborated.

- the evaluator may discover that the original evaluation design cannot be fully carried out within the time required. This can happen even if the evaluation is fairly well planned. Alternatively, the evaluator may want to suggest that the original design is changed so that more time can be allowed to examine features of the programme which were not part of the original design.

- once the evaluation is underway, the evaluator may encounter resistance from programme administrators, programme beneficiaries or other stakeholders. For example, they may refuse to make data available.

The management structure needs to be aware of the potential for such problems to arise once an evaluation is underway.

3.2. Elaborating the evaluation project

The evaluation project is a sequence of logical steps, starting out from the formulation of problems and interests motivating the evaluation and arriving at a series of questions that can be addressed in an analytically acceptable way.

The seven steps involved in elaborating an evaluation project are as follows:

- identify the goals of the evaluation;
- delineate the scope of the evaluation;
- draw up the analytical agenda;
- set benchmarks;
- take stock of available information;
- map out the work plan; and
- select the evaluator.

These seven steps should be gone through in virtually all evaluations. We will now discuss each of them in turn.

3.2.1. Identifying the goals of the evaluation

The question to be asked before any other when preparing an evaluation undertaking is: ***why?*** For what purposes are we launching the evaluation? The

answers to this first question will have a strong bearing on the replies to the subsequent ones.

Evaluations often have to be carried out because of an obligation laid down in the programme's legal base, stipulating typically that a report should be available prior to the expiry of the programme.

Since the adoption of the Communication on Evaluation (on 8 May 1996) in the framework of the SEM 2000 initiative, the general rule has been introduced that a proposal to renew a multi-annual programme has to build on the results of an evaluation of its achievements to date. Operational expenditure outside a multi-annual framework has to be reviewed at least every six years.

As mentioned in Chapter 1, there are three *specific* reasons why programme evaluations are conducted:

- to improve management;
- for reasons of accountability; and
- to assist in the allocation of budgetary resources.

The contents of the evaluation and the style of the report will differ according to where the relative emphasis is put between these elements,. If stress is laid on **improving management**, screening the implementation and delivery mechanisms of the programme will occupy a central position in the evaluation. The report can stay quite technical as its customers are mainly the Commission services, intermediary offices and direct beneficiaries.

If **accountability** is at the forefront, the evaluation is likely to focus on the effectiveness of the programme as reflected by empirical evidence and in the perception of the main stakeholders, as well as on possible side-effects and specific issues associated with, for example, equity and transparency. The style should take into account the fact that the wider audience may lack the specialist vocabulary and detailed technical knowledge associated with the programme.

If, as in the case of evaluations springing from SEM 2000 obligations, part of the emphasis is on **programme renewal and related budgetary needs**, the goal of the evaluation should be, among other things, to shed light on the cost-effectiveness of the programme, its continued relevance, and (possibly) a comparative analysis of alternatives. Here again, the style should ensure good legibility for decision-makers and opinion-formers.

The goals of an evaluation should, of course, be **realistically attainable**. For instance, consider the case of a 4-year programme which is in its first generation. Taking into account the time normally required for the legislative authority to decide on the adoption of a new proposal, the evaluation report should, in principle, be available in the second half of the third year. This implies that the evaluation should be launched no later than early in the first half of that year. At that time, however, it will only be possible to make observations on two years of programme execution, at best, which will permit little more than a

progress report containing an assessment of outputs and very preliminary indications on results. On the other hand, programmes in their second generation or beyond should be evaluable in terms of results and outcomes, and thus evaluations should be able to address key issues such as effectiveness.

Other factors can affect the attainability of an evaluation's goals, for instance its budget. In addition, in some cases, the controversy surrounding a programme can be such that as soon as fundamental matters are broached, the risk of the evaluation becoming entangled in the political cross-fire may become large. This would reduce the evaluation's credibility. In such cases, it may be wiser to scale down the ambitions of the evaluation.

3.2.2. Delineating the scope of the evaluation

Delineating the scope of an evaluation means asking the question: ***what is to be evaluated?*** Regardless of how comprehensive one intends an evaluation to be, delineating its scope is a vital part of the evaluation project. It would simply be an endless task to look into every imaginable facet of a programme, or into all its actual or potential ramifications with other programmes at a European or national level. For instance, if one were to evaluate, in a truly comprehensive manner, EU policy on the development of rural areas, one should not only evaluate the effects of Objective 5b expenditure under the Structural Funds, but also the impact on rural areas of the entire set of European policies as well as the interaction of these policies with those at the national level.

Typically, an evaluation's field of investigation, particularly the part to be dealt with in depth, has to be circumscribed from an institutional (EU vs. national or local level), temporal (period reviewed) and geographical (part of the EU territory) point of view.

A second major issue of scope, related to the comments above on the goals of an evaluation, is which key evaluation issues one plans to observe and measure. As explained in section 2.2.2., these are *relevance*, *efficiency*, *effectiveness*, *utility* and *sustainability*. Apart from the reply given to the previous question on goals, this choice will be influenced by factors such as the availability of data, time constraints and limitations on financial resources.

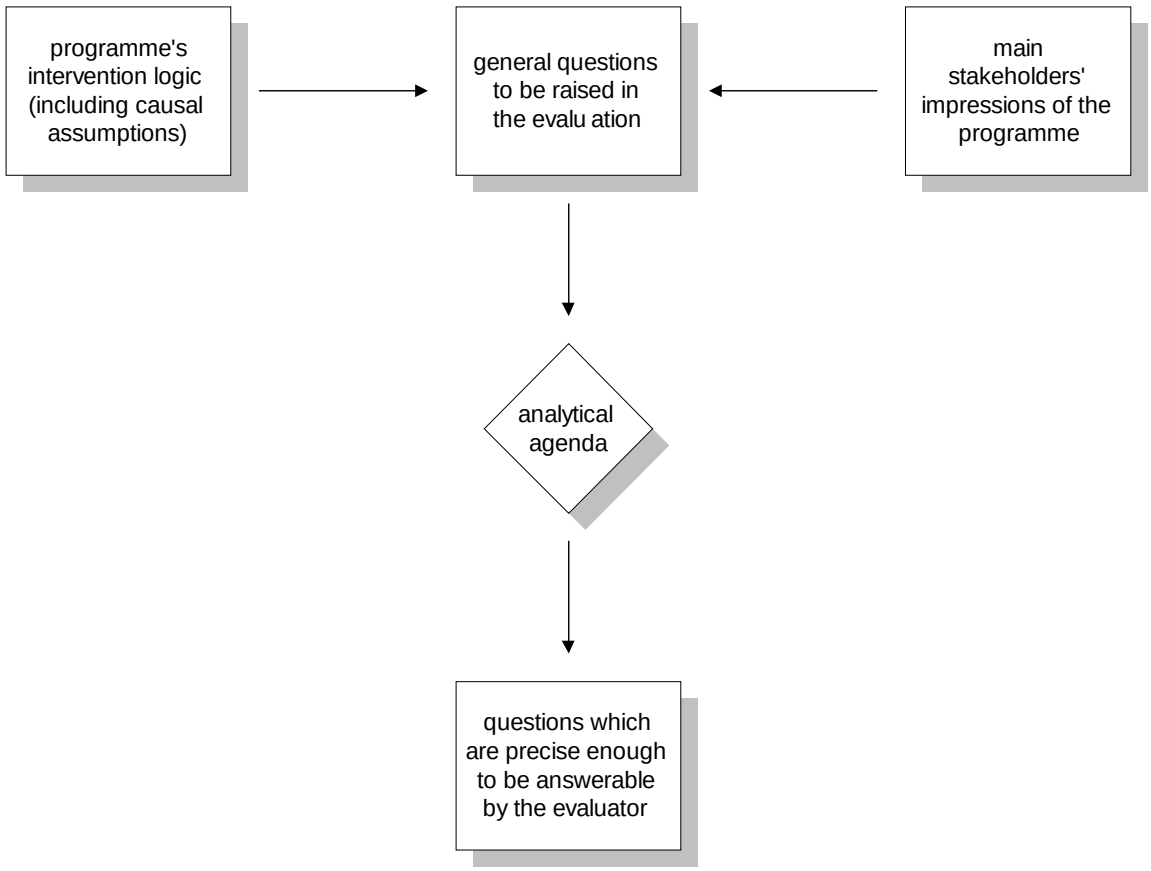
A key feature of the scope of an evaluation which seeks to provide lessons for the future of a programme and its management is that it examines, with the benefit of hindsight, the validity of the *intervention logic* (see section 2.2.1. above) as formulated when the programme was launched. A central question to be asked is: *has the causal link from inputs* (financial and human resources) *to outputs* (goods and services produced by the programme), *and subsequently to the achievement of results and outcomes occurred as initially envisaged, and if not why?* We will discuss this in more detail below.

3.2.3. Formulating the analytical agenda

Once there is a clear idea about what purposes an evaluation should serve and what major issues it should address, the next step in the preparation of an evaluation is to draw up the ***analytical agenda***. This is a *logical structure imposed on the different questions to be asked in the evaluation.*

The aim of an analytical agenda is to transform the general, often vague, questions which those calling for the evaluation have in mind into questions which are precise enough to be manageable by evaluation research methods (based, invariably, on research methods derived from economics, social sciences, management science, and so on). This process is shown in Figure 3.1. below.

Figure 3.1. The process of formulating an analytical agenda



The analytical agenda is simply a way of transforming the *general* questions into more *precise* questions. The two main sources for the general questions are:

- the programme's original intervention logic, i.e. the "theory" of what it is supposed to achieve and how it is supposed to achieve it; and

- the impressions of the main stakeholders.

In relation to the intervention logic, special attention needs to be paid to the *causal assumptions* which are usually hidden beneath the surface. The most important assumptions relate to *how* the programme is supposed to generate its

supposed effects, and the state of the programme's external environment (i.e. *how* it relates to other policy interventions and other external factors).

Retrieving the original intervention logic of a programme is sometimes easier said than done. Official documents often do not contain any systematic description of causal assumptions. Even the programme's objectives may only be stated in a very limited fashion. Furthermore, the *collective memory* of Commission services may not be that long (e.g. because of a high turnover of programme officials). Often, substantial documentary research may well be needed in order to retrieve the correct interpretation of official goals. In any event, when a programme's objectives are not given a sufficiently transparent and precise meaning, it can be very difficult to judge its success.

Where the general and specific objectives of a programme have to be reconstructed from scratch, this should be done transparently by the management structure, preferably under the responsibility of a steering group.

A second useful input into the process of drawing up the analytical agenda is to collect and present the main stakeholders' impressions about the programme (its successes, failures, evolving context etc.). These should then be examined critically as "working hypotheses" in the evaluation. This process will both enrich the analytical agenda for the evaluation and reinforce its focus on utility. It should not, however, prejudge the conclusions to be reached by the evaluation.

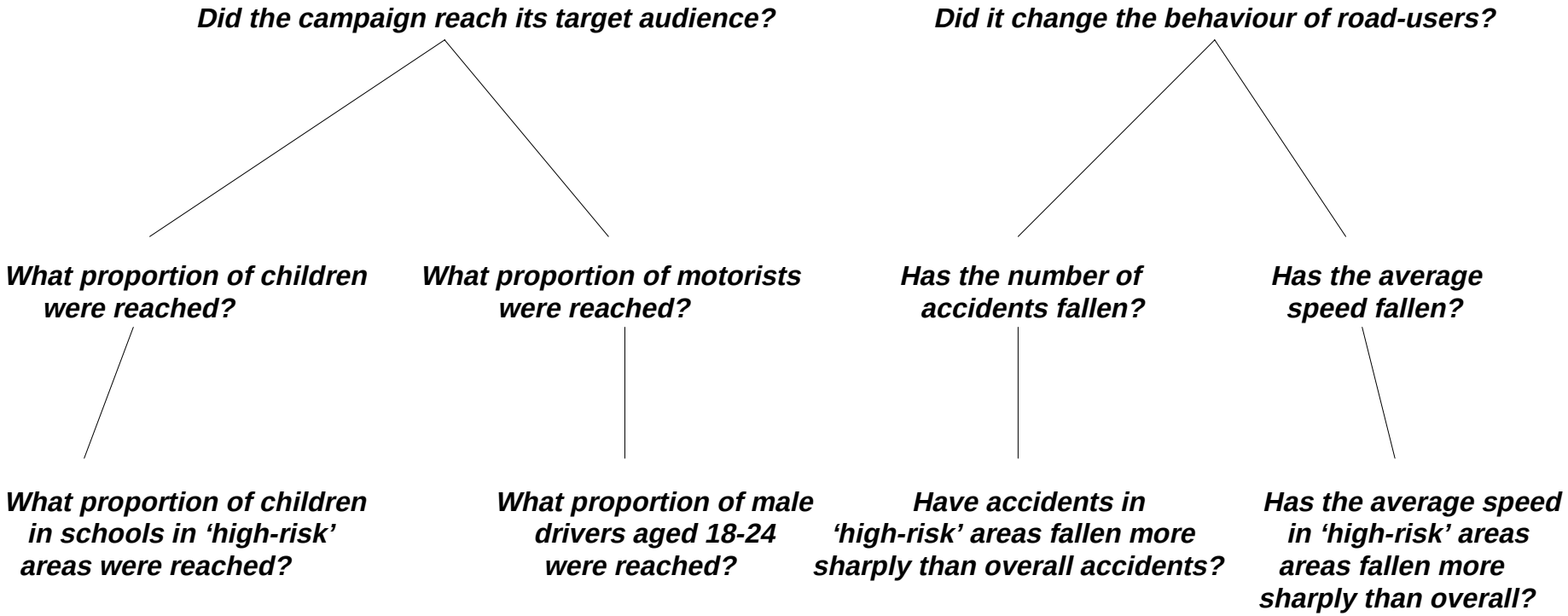
Once the general questions have been identified, the analytical agenda should be drawn up. Essentially, this means arriving at a set of precise questions which can be answered by an evaluator using accepted research methods. The analytical agenda imposes a logical structure on the various questions which should be addressed. The simplest logical structure for our purposes is a *hierarchy*.

At the lowest level of this hierarchy are the most detailed and refined questions. These are sufficiently precise and well-defined to be answerable by accepted research methods. As one moves up the hierarchy, it should be clear that the more detailed questions towards the bottom provide a basis for investigating the more general questions towards the top. A simple example of such a hierarchy is given in Figure 3.2., which is based on a project-based evaluation of a publicity campaign designed to raise awareness of road-safety in a medium-sized town.

The campaign was directed at the general public, but with particular attention to school children, especially those in 'high-risk' areas and male drivers aged 18-24. The evaluator was asked to find out who the campaign reached and whether it changed the behaviour of motorists. For the sake of simplicity, we have chosen an example involving the evaluation of a *project*, the same sort of principles apply to establishing analytical agendas for the evaluation of a *programme*.

Figure 3.2. An analytical agenda imposes a hierarchy on the questions raised in an evaluation

Example: a publicity campaign designed to promote awareness of road-safety



Once the analytical agenda for the evaluation has been drawn up, **those responsible for commissioning the evaluation need to ask whether the programme is indeed *evaluable***. The questions that were identified when the analytical agenda was drawn up should be answerable by an evaluator using appropriate research methodologies. **To know whether the questions can be answered with an acceptable degree of credibility, it is often advisable to perform an *evaluability assessment***. If a programme is not evaluable (e.g. because adequate data are not yet available), this can lead to a decision to postpone the evaluation or to draw up a new, more realistic analytical agenda. Nevertheless, it should always be remembered that **it is better to have imprecise answers to important questions than to have precise answers to unimportant questions**. Thus, even if a programme is only partially evaluable in terms of a given analytical agenda, it may still be useful to proceed with the evaluation.

3.2.4. Setting benchmarks

Evaluation is about revealing the 'value' of a programme. This involves making value judgements on the degree to which a programme's performance has been 'good' or 'bad'. **Predetermined and transparent benchmarks are needed to ensure that value judgements do not become arbitrary**.

What are the criteria by which to rate the observed effects of a programme? What standards should be used to pronounce on the proper functioning or success of a programme? An obvious place to start would be the programme's objectives as expressed by expected outputs, results, and outcomes. However, setting benchmarks may prove difficult for a number of reasons:

- objectives can sometimes be expressed in very vague terms.

- a single programme may have multiple objectives, either in terms of results or outcomes, some of which may carry relatively more weight, or even be incompatible with others.

- objectives may also evolve over time, as the programme's environment evolves. A notable example is the Phare programme for assistance to the associated countries of Central and Eastern Europe, whose goals have undergone substantial modifications since it was first introduced.

There is, however, more to benchmarking than simply reconstructing, clarifying and ordering objectives. **Benchmarks should ideally allow us to compare the programme's performance with that of other policy instruments in the same field of action or in a related one.** This is important because **if a programme falls short of achieving its objectives, its performance may not necessarily be unsatisfactory**. It may compare favourably with results achieved by similar programmes executed in the past, or by national or local governments, or countries outside the Union. A comparative perspective may suggest that the ambitions for the programme were unrealistically high, rather than that the programme itself has failed.

In principle, then, there are three different axes on which benchmarks which can be established:

- *time* - benchmarks which compare the same programme over time (to what extent are the programme's objectives being met this year compared to last year?);

- *space* - benchmarks which compare the same programme in different areas (to what extent are the programme's objectives being met in one area compared to another?). and

- *time and space* - benchmarks which compare the programme with other policy instruments which are roughly similar.

When judging programme performance by means of benchmarks, the fundamental caveat needs to be kept in mind that benchmarks may have been reached by virtue of developments that are not attributable to the programme. **An evaluation should try to separate out these developments, in order to identify the *net* effect of a programme on the achievement of its objectives.** Data on the respect of benchmarks have to be interpreted carefully. This is particularly true of objectives that can be influenced by a whole range of exogenous factors, such as national policies on which the EU programme may have little or no effect. The issue of *net attribution* is of fundamental importance in the choice of evaluation design, as will be discussed in Chapter 4.

3.2.5 Taking stock of available information

The next step in the preparation of an evaluation project is to take stock of available information. **For most programmes, the monitoring system should be the first source of existing information**. The quality of the monitoring data will be an important determinant of the success of the evaluation. However, monitoring data will rarely be sufficient. Other existing material which is readily accessible can include professional literature, media articles, administrative data or published statistics.

It is often helpful to produce a *research synthesis* of the current state of knowledge about a problem and about remedies through policy intervention and public expenditure. This can serve to guide the evaluation's analysis and choice of method, especially with respect to questions on relevance and effectiveness.

Clearly, a programme based on a sound ex ante evaluation will have taken account of the existing knowledge at the time of its inception. However, not all EU programmes have benefited from such a systematic inquiry in the past and, even if they have, several years may have elapsed since the programme was launched, calling for an update of the research synthesis.

By listing the information that is available and comparing it with the needs ensuing from the analytical agenda, the inventory will point to the principal information gaps which, in turn, set the data collection and interpretation tasks to be undertaken by the evaluation. However, it is important to proceed with caution. The analytical agenda may be the result of a maximalist approach,

raising questions on which data can only be of doubtful quality or obtained at large cost. Some of these questions may be fairly remote from the key objectives of the programme. Evaluations face a time and budget constraint, so that before launching data collection activities, it should be decided whether the data to be generated are liable to shed any significant new light on the subject under scrutiny. It should also be remembered that an evaluator can always turn to existing literature as a source of data when conducting the evaluation. If a literature review is foreseen as a potential data collection technique, it may not be necessary to conduct a research synthesis as well.

3.2.6. Mapping out a work plan

Once the previous steps have been completed, it should be possible to draft a work plan which sets out the investigations that need to be conducted by the evaluation, taking into account the chief questions raised by the analytical agenda and the information gaps which have been identified.

These investigations should be described in sufficient detail to provide a good, albeit provisional, picture of the data collection and analysis tasks lying ahead, and, where possible, of the methodologies to be employed.

In order to keep them manageable, it often proves useful to divide the various tasks into different stages and to set a corresponding time-table for the delivery of the consecutive evaluation parts (e.g. interim reports).

The work plan is also the appropriate place for costing the evaluation and its components. When the evaluation is done internally, an estimate should be given of the global amount of time to be spent by officials, as well as of other administrative expenditure. When external experts are relied on, estimates should be made before putting out the call for tenders. This is done in order to verify that the budget set aside for an ex post or intermediate evaluation by outside experts is compatible with the ambition of the analytical agenda contained in the work plan. The Communication on Evaluation of 8 May 1996 specifies that the overall budget for all evaluation activities throughout the life time of a programme might amount to up to 0.5% of the programme's budget.

Costing should always be realistic. All too often, evaluations arrive too late or do not achieve what they set out to do because initial expectations were too high. For example, if one wants to engage in a serious activity of data collection that cannot be connected to a monitoring system, this can be quite expensive. Also, time and money are only very partial substitutes. Raising the budget may cut the time otherwise needed, but usually the relationship between these two factors is not simple.

3.2.7. Selecting the evaluator

Drawing up the analytical agenda and mapping out the work plan are extremely useful exercises to perform before selecting the evaluator. In particular, once it is clear what type of questions the evaluation needs to ask, and once its budget and time-schedule have been determined, it should be easier to decide between internal and external evaluation.

Evaluation tasks vary widely and the choice of evaluator should reflect this. Some evaluation activities are technically highly complicated, costly, and of such long duration that they require the dedicated participation of highly trained specialists. On the other hand, there are many evaluation tasks that can be easily understood and be carried out by persons without any strong sector-specific background. Indeed, some professional distance from the subject to be examined can often be an advantage, since it may allow the evaluator to take a more objective and independent view of a programme.

The technical capacity of an evaluator is an important selection criterion, but it is not sufficient by itself. Other important issues in selecting evaluators include:

- the ability to obtain access to information and actors
- knowledge and previous experience of the programme area
- independence of the evaluator from the main stakeholders
- specific characteristics associated with the policy area (e.g. the evaluator may be required to work in hazardous areas)

If a decision is made to appoint an external evaluator, it should be noted that there are a number of different types of organisation which can perform an external evaluation. Two of the most often used are:

- **management consultancies** - these vary from large, multinational firms which have considerable experience in carrying out a range of different evaluations to smaller firms which possess a narrower, highly subject-specific expertise. Such firms are often perceived by stakeholders to embody a "businesslike" approach (although in some public sector contexts, this can be a disadvantage). Typically, such organisations can perform evaluations relatively quickly and tend to possess excellent presentational skills. However, they may also have disadvantages. Firstly, their prices may be relatively high compared to other types of organisation. Where their prices are competitive, this may be a deliberate attempt to win more work by "loss-leading". Alternatively, they may seek to reduce their own costs by applying a pre-packaged solution to a given evaluation problem rather than specifically tailoring an evaluation to the needs of those sponsoring it and of the other principal stakeholders. Finally, *a risk with management consultancies is that they may promise an evaluation but deliver an audit*.

- **academic institutions** - academic experts are likely to offer a high degree of methodological expertise in evaluations. Some may also possess a high degree of subject specific knowledge. Stakeholders may tend to perceive academics as being relatively independent, and this can be an advantage in circumstances where a management consultancy might be viewed with suspicion. An academic institution may represent better value-for-money compared to a management consultancy, but can often be less flexible. Finally, *a risk with academic institutions is that they may promise an evaluation but deliver a scientific study*.

For large programmes, or programmes with a varying regional impact, it is often helpful to rely on a *consortium of evaluators*. This allows for a combination of different types of evaluation organisation to be used. Normally, one organisation

will be chosen to supervise the overall evaluation and produce a synthesis report. Individual aspects of the programme (or different regions) will then be divided between the different members of the consortium.

There are several criteria which the ideal evaluator should satisfy: specialist knowledge of the particular field, expertise in evaluation, independence and external legitimacy, ability to work to required deadlines, value-for-money and integrity. Of course, no one can fully satisfy each of these criteria. In the real world, choosing an evaluator necessarily involves compromising on one or more points.

3.3. Drawing up the terms of reference

Clearly defined *terms of reference* are vitally important where an evaluation is to be conducted by an external expert, and can also be of tremendous use when it is to be performed in-house. **The terms of reference outline the work to be carried out by the evaluator, the questions to be dealt with and the time schedule. They allow the sponsors of the evaluation to define their requirements and allow the evaluator to understand clearly what is expected of the work to be undertaken**.

The terms of reference must reflect the specific circumstances of the programme being evaluated. In the case of evaluations entrusted to external contractors, the terms of reference attached to the contract may differ from those initially prepared at the call-for-tenders stage, following discussions and negotiations with the chosen contractor, who may bring his knowledge and experience to bear. In this case, it is important that potential evaluators know what scope they have to refine the original evaluation project before the contract and final terms of reference are finalised.

Some of the main elements which would normally be covered in the terms of reference are:

- the legal base and motivation for the evaluation
- the future uses and users of the evaluation
- a description of the programme to be evaluated
- the scope of the evaluation
- the main evaluation questions
- the methodologies to be followed in data collection and analysis
- the work plan, organisational structure and budget
- the selection criteria for external evaluators
- the expected structure of the final evaluation report

We will briefly discuss each of these in turn.

3.3.1. The legal base and motivation for the evaluation

It is helpful for both the evaluator and the sponsors if the terms of reference specify the legal and contractual requirements upon which the evaluation will be based.

3.3.2. The uses and users of the evaluation

Evaluators need to know how the findings of the evaluation will be put to use, who are the primary intended users and what results are expected of the evaluation. Answers to these questions will help the evaluator to identify the main purposes which the sponsors have in mind in commissioning the evaluation. These purposes will, in turn, affect the specific questions which the evaluator should address in his work, the relative emphasis on programme implementation and outcomes and the level of programme detail at which he will seek to provide answers.

3.3.3. The description of the programme to be evaluated

The terms of reference should normally include a succinct but comprehensive definition of the programme to be evaluated (including, for example, its intended target population, its general and specific objectives, its inputs and outputs, and its delivery systems).

3.3.4. The scope of the evaluation

The terms of reference should specify what part of the programme the evaluation should cover and what aspects of the programme are to be considered. At this stage, you can refer to the evaluation project elaborated above (see section 3.2.2. in particular).

Some of the important questions to ask when defining the scope of the evaluation are as follows:

- Will the evaluation be expected to cover the entire programme under consideration? If not, the terms of reference should indicate clearly which part of the programme is to be excluded (proportion of the budget, geographical areas, periods of time, or specific aspects, activities or client groups).

- Is the programme to be evaluated in isolation, or is the evaluator expected to examine links between it and other EU programmes?

- Should the evaluator examine the extent to which the programme's *expected* outputs, results and outcomes have been realised (i.e. the extent to which specific and general objectives have been achieved)? Are unforeseen results and outcomes, whether negative or positive, to be examined as well?

3.3.5. The main evaluation questions

It is important to specify the evaluation questions from the analytical agenda (as explained in section 3.2.3. above) in order to provide the evaluator with precise guidelines as to the exact information needs of sponsors and stakeholders. These information needs will tend to differ according to whether a *formative evaluation* or a *summative evaluation* is to be conducted.

Of course, one of the main questions to be addressed in most evaluations is whether the programme's *intervention logic* is valid. It will be remembered that the intervention logic describes how the programme's inputs (human and financial resources) have been converted into outputs (goods and services

produced by the programme), and how this in turn leads to the attainment of results and outcomes.

3.3.6. The methodologies to be followed in data collection and analysis

When drawing up the terms of reference, the sponsors will normally wish to give clear guidance on the data collection and analysis methods to be followed by the evaluator. Whilst it is important to note that both internal and external evaluators are likely to benefit from such guidance, it should also be remembered that there is no single, universally applicable methodology.

The methodology to be used for data collection and analysis must be appropriate given the specific circumstances of the programme to be evaluated and the particular questions to be addressed. In the case of external evaluations, broad guidelines can sometimes be preferable, at least at the call for tenders stage. This allows the chosen contractor to use any knowledge and experience to refine the suggested approach through a process of negotiation and discussion with the sponsors. The final terms of reference, as attached to the contract, can then be more precise.

3.3.7. The work plan, organizational structure and budget

The *work plan* for the evaluation should include factors such as the length of the contract and the deadline for reporting. It may also be appropriate to provide the evaluator with guidance on existing data sources (e.g. monitoring data) and relevant contacts to be made.

Specifying the evaluation's *organisational structure* involves delineating the role of different actors (especially important if the evaluation task is to be divided among different evaluators - for example, between internal and external evaluators); establishing reporting responsibilities (including, where appropriate, contact with evaluation steering groups, programme managers, other Commission services and Member State administrations); and identifying the procedure to be followed to disseminate and use evaluation results.

Except where the evaluation is to be conducted wholly internally, the *budget* for the evaluation should also be stated, including per diem expenses and reimbursable travel costs.

3.3.8. The structure of the final evaluation report

There is no universally acceptable structure for evaluation reports, although all reports should include an executive summary as well as a copy of the terms of reference (usually as an appendix). A typical evaluation report structure is presented in section 5.2.1.

Where to look for more information

The interested reader can consult a variety of sources on preparing and managing evaluations, including Conseil scientifique de l'évaluation (1996). MEANS Handbook Number 1 on *Organising Intermediate Evaluation in the context of Partnerships* is specifically designed to be used in the case of EU Structural Funds. However, it contains much that can be applied to other areas of EU activities. There is also a typical example of a terms of reference prepared by C3E.

4. Conducting evaluations

Conducting an evaluation involves choosing a particular *evaluation design*, which is a framework for describing a programme and testing hypotheses about its effects.

A given evaluation design allows an evaluator to choose one or more *data collection techniques*. These are the methods used to gather information about a programme. Evaluation designs also lead to a choice of *data analysis techniques*. These are the methods used to interpret the information which has been gathered.

At the outset, it is worth highlighting the *golden rule* about evaluation techniques:

Golden rule: there are **no** golden rules.

In other words, there is no single evaluation methodology which is universally applicable. Instead, **the choice of techniques should be determined by the particular evaluation problems at hand**.

- Poor evaluations often result from an arbitrary choice of method at the beginning of the undertaking (e.g. based on whatever data are immediately available) which then proves to be unsuitable later on.

- The best evaluations use proven techniques for collecting and analysing data, and the choice of techniques is justified in relation to the problems posed by the particular evaluation. They often employ more than one technique, so that the strengths of one can balance the weaknesses of another, hopefully giving rise to complementary findings.

In this chapter, we will introduce the concept of evaluation designs and show how it has an important role in determining the credibility and analytical rigour of an evaluation. We will then discuss a number of techniques for data collection and analysis that can be used in different evaluation designs. The current guide cannot provide a complete description of all the different analytical techniques from the fields of statistics, economics or the social sciences. Instead, it gives a overview of the basics of evaluation research that are worth keeping in mind when actually conducting evaluations.

4.1. Introducing evaluation designs

An ***evaluation design*** is a *model which is used to describe a programme and provide evidence on the effects which may be attributable to it*. Evaluation designs are of central importance in examining the validity of the programme's *intervention logic*, i.e. the theory of how the programme achieves its objectives by generating certain effects. In this section, we will discuss some of the main features of evaluation designs.

For the sake of simplicity, we will assume that a programme can be evaluated through a single evaluation design. For many EU programmes, this is obviously not the case. Very often, programmes have a diverse range of effects (and often there are sub-programmes or large projects which need to be evaluated separately). In reality, therefore, some combination of evaluation designs may have to be chosen.

For the sake of presentational clarity, we will start with a discussion of the ideal experimental design, which is essentially a theoretical construct. As will be seen, in the real world there is no such thing as the ideal experiment. We will therefore proceed to discuss the threats to causal inference which can arise in the real world, and then move on to describe various real world evaluation designs. The evaluation designs available in the real world are divided into two approaches. The first is based on attempting to attribute causality, i.e. designs which enable us to say whether or not observed effects were caused by a programme. The second is based on describing the programme and its supposed effects.

4.1.1. Causality and the ideal experimental design

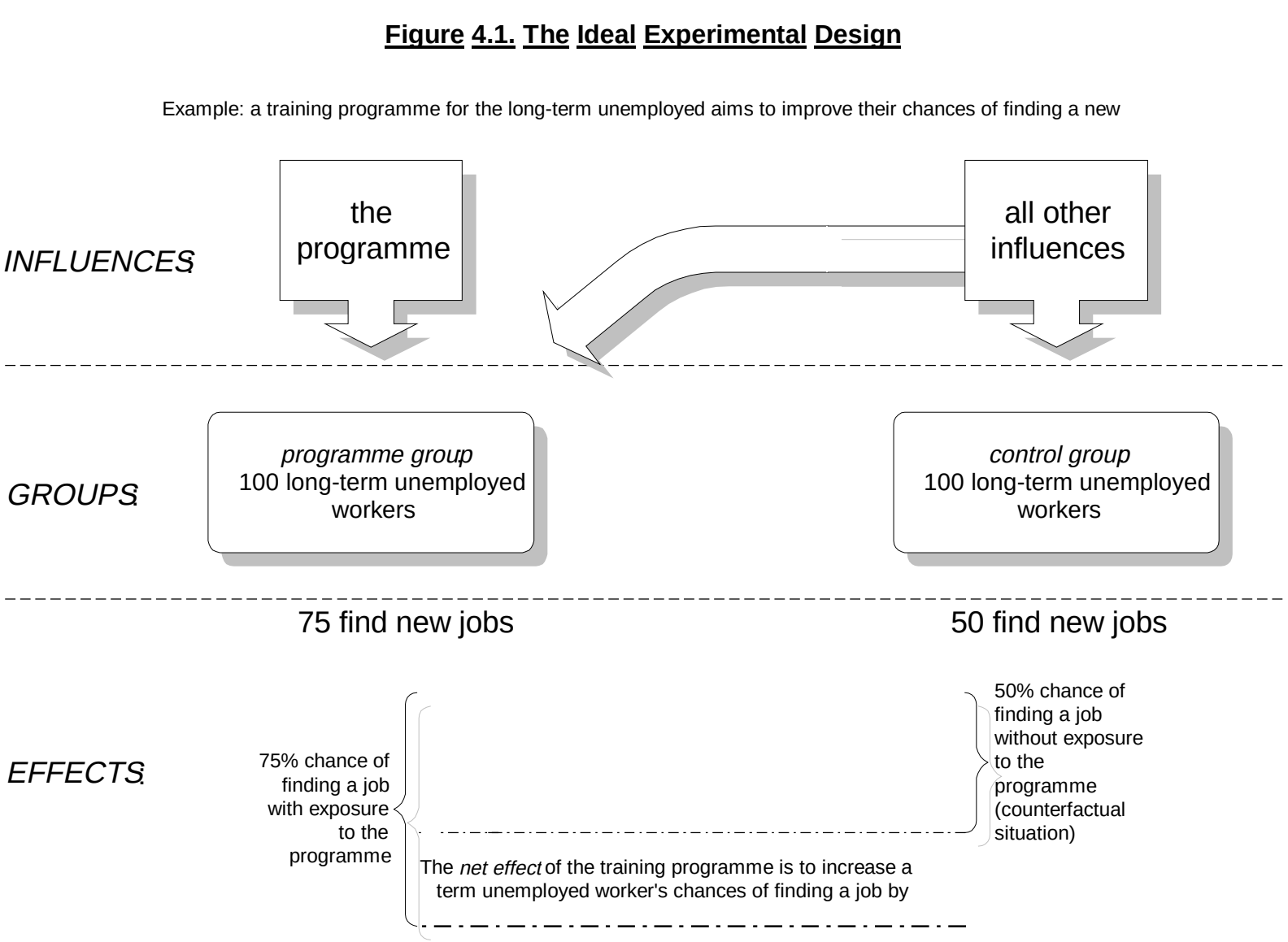
It should be clear that, since evaluation designs help us to investigate the effects which may be attributable to a programme, they are closely related to the concept of *causality*.

Let us recall the example of the local road safety awareness campaign discussed in the previous chapter. Suppose there is a fall in road accidents in the local area after the campaign. Is this benefit definitely attributable to the campaign itself? The campaign may have been launched at the same time that a national reduction in the speed limit for motor vehicles was introduced. Or, suppose that accidents in the local area actually rise after the campaign. Does this mean that it did not have a beneficial effect? Not necessarily, since accidents may have risen by even more without the campaign. Similarly, if there is no change in the number of accidents after the campaign, this may mean that the campaign succeeded in arresting the rise in road accidents.

The existence of a programme may be a *necessary condition* for the resulting effects to occur, but it may not be a *sufficient condition*. For example, the evaluator of the road safety awareness programme may indeed find that without the programme, there would not have been a fall in local road accidents. However, it may also be true that certain other factors (e.g. local road conditions, a fairly young average age for motor vehicles, etc.) are also required in order to reproduce the resulting effects. **Alternatively, a programme may be *sufficient* but not *necessary***. In the road safety awareness programme referred to above, the evaluator may find that the fall in local road accidents after the campaign would have happened anyway, e.g. because of the introduction of a new national speed limit or due to improved weather conditions affecting local roads. **Finally, the programme may be neither *necessary* nor *sufficient*.** The observed effects may simply have nothing to do with the programme.

When we say that certain effects were *produced* or *caused* by a programme, this means that if the programme had not been there or had been there in a different form or degree, those effects would not have occurred, or would not have occurred at the same level. This means that it is important to have a clear idea of what would have happened without the programme. This is called the ***counterfactual situation***.

Figure 4.1. The Ideal Experimental Design



Ideally, we want to derive the counterfactual situation with certainty. We could do this by comparing two groups which are identical in all respects except that one group (which we will call *the programme group*) is exposed to the programme whilst the other group (which we will call *the control group*) is not. An illustration of such an *ideal experimental design* is given in Figure 4.1. above.

In this example, we have a training programme designed to increase the chances of long-term unemployed workers finding new jobs. Two hundred long-term unemployed workers who have the same skills and experience are divided between the programme group and the control group. The 100 members of the programme group are exposed to the training programme, whilst the 100 members of the control group are not. The groups are identical in all other respects and both groups are exposed to all other influences apart from the programme.

After the programme, 50 workers in the control group have found new jobs. This is our estimate of the counterfactual situation - without the programme, there is a 50% chance of an unemployed worker finding a new job. However, amongst the programme group, 75 workers have found a new job. Therefore, we might conclude that the net effect of the programme is to increase a long-term unemployed worker's chances of finding a job by half.

In the real world, however, this ideal experiment does not exist since we can never be absolutely certain that the programme group and the control group are identical in all respects except for exposure to the programme. The two groups are, after all, made up of different members and will be therefore be different in some ways even if these differences do not show up in average measures.

The potential non-equivalence of the two groups means that the counterfactual situation has to be estimated rather than derived. This immediately weakens the validity of any *causal inference* about the programme. In other words, there are plausible alternatives which may explain the effects which would otherwise be attributed to the programme itself.

Plausible alternatives pose problems to causal inference. The evaluator's task is to try to overcome these problems by choosing an evaluation design which is robust to them. We will see how avoiding different types of problem helps to determine the choice of evaluation design in the real world. To do this, we must first examine the threats to causal inference in more detail.

4.1.2. Threats to causal inference

In the real world, where there is no such thing as the ideal experiment described above and where there are potential threats to the validity of any causal inference, we need some means of choosing between different evaluation designs. In selecting an evaluation design, the main criteria which we consider are internal and external validity.

Internal validity *refers to the confidence one can have in one's conclusions about what the programme actually did accomplish*. A threat to internal validity implies that the causal link between the programme and the observed effects is uncertain due to some weakness in the evaluation design. It may be thought of

as questions of the following nature: what confidence can one have in the estimate of the counterfactual situation? Could not something else besides the programme account for the observed effects? For example, how certain can one be of the contribution of programmes to promote the use of alternative energy sources to the rise in the share of these sources in total energy consumption?

External validity *refers to the confidence one can have about whether or not one's conclusions about the programme can be generalised to fit circumstances, times, people, etc. other than those of the programme itself*. A threat to external validity is an objection that the evaluation design does not allow causal inference about the programme to be generalised to different times, places or subjects to those examined in the evaluation. For example, when conducting an evaluation of support to small and medium-sized enterprises in Saarland, to what extent can conclusions be carried over to other regions, such as Bavaria, Picardy, Umbria or Andalucia?

External validity is of central concern where case studies are used and is also of paramount importance in the evaluation of pilot actions. It should always be a standard consideration in determining the scope of an evaluation (see 3.2.2. above).

Evaluators must ask themselves what sort of decisions are likely to be made as a result of an evaluation, and be aware of the corresponding challenges to internal or external validity.

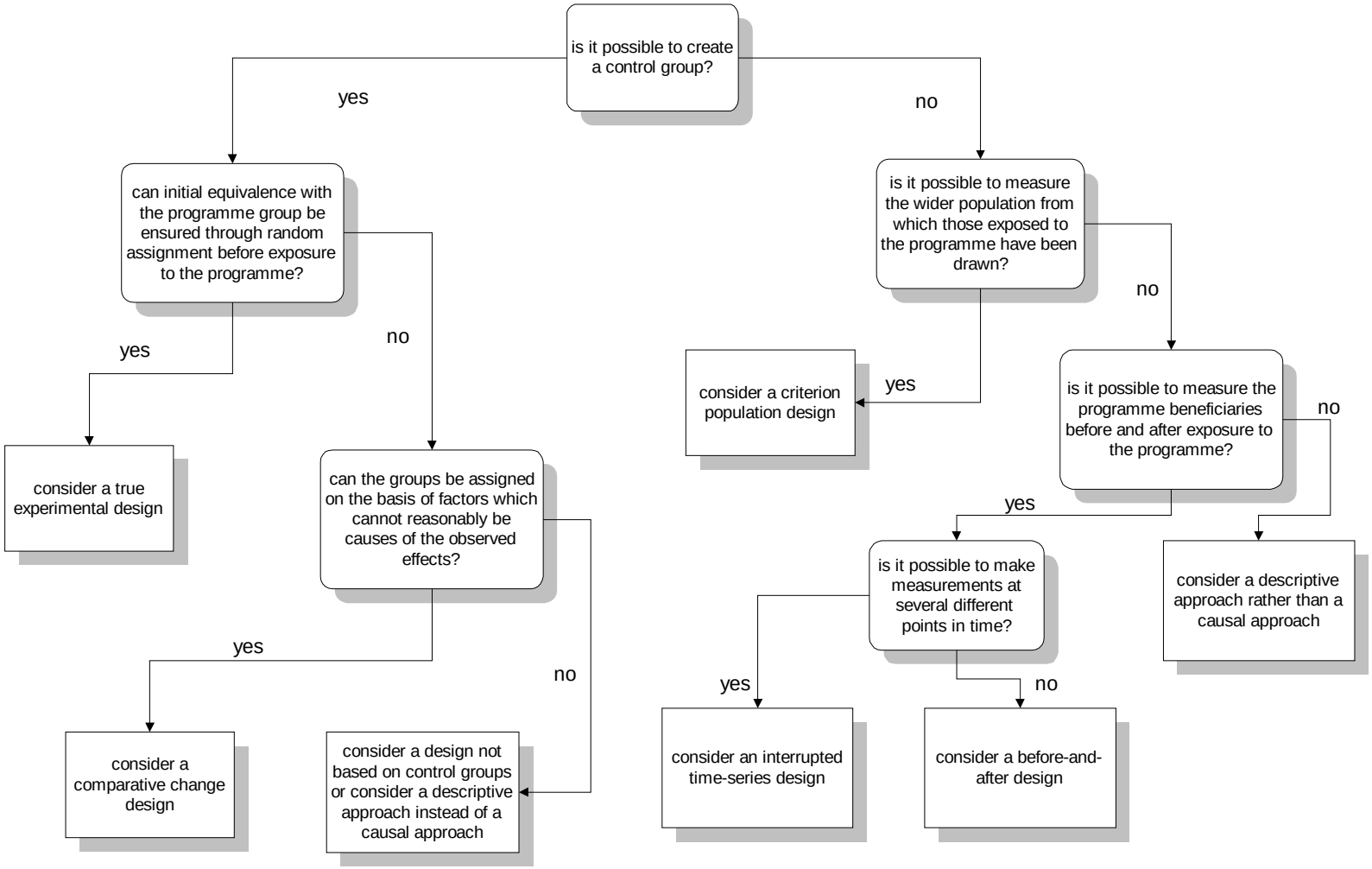
4.1.3. The causal approach to evaluation designs

We are now in a position to briefly examine some of the main evaluation designs which are available in the real world. In this section, we will cover those evaluation designs which can be used to allow the evaluator to engage in causal inference even though the conditions of the ideal experiment can never be reproduced. In the following section, we will examine those designs which are appropriate to situations where the evaluator seeks to provide a description of the programme and its supposed effects.

Evaluation designs in the causal approach attempt to estimate the counterfactual situation in some way, rather than deriving it as is the case in the ideal experiment. A helpful way of classifying the various causal designs is to ask whether the estimate of the counterfactual situation is derived from (i) the same subjects at one or more previous time periods; or (ii) a group of comparable subjects, i.e. a control group. Figure 4.2. below shows the selection criteria which can be used to choose between each of the different designs which we will discuss.

One approach which is based on the use of control groups is provided by ***true experimental designs***. True experiments are the best real world approximation to the ideal experimental design. Recognising that there is the problem of the potential non-equivalence of the programme group and the control group, true experimental designs attempt to ensure the initial equivalence of the two groups by creating them through some *random* process (e.g. by picking names out of a hat).

Figure 4.2. Criteria for selecting an evaluation design (the causal approach)



Causal inference is usually strong under such designs, since most of the factors that determine effects other than the programme should be evenly distributed between the two groups - they have, after all, been *randomly* assigned. In practice, however, such designs may often be extremely difficult to arrange and implement. More specifically, the evaluator is very rarely in a situation where he himself can determine, before the programme starts, who is to be exposed to it and who is not. Thus, it is hardly possible to apply true experimental designs to evaluating, for example, the effects of scholarships awarded under the ERASMUS programme on the careers and attitudes of beneficiaries, since students are not selected randomly for inclusion in the programme.

A more practical approach is available through the use of ***quasi-experimental designs***. Control groups can still be used, but these have to be created through some non-random process. Alternatively, one can examine programme beneficiaries before and after their exposure to the programme. The first quasi-experimental design which we will look at is, in fact, called the *before-and-after design*. In this approach, one simply compares the situation after the programme with the situation beforehand and attributes any difference to the programme. Administering a before-and-after evaluation design is relatively easy, but causal inference tends to be quite weak. There is always the possibility that something else besides the programme may account for all or part of the observed change over time.

An improvement on the before-and-after design is the *interrupted time-series design*. As can be seen from Figure 4.2., it involves obtaining additional information over time both before and after exposure to a programme in order to create a time-series of observations. In principle, we should have more confidence in claiming that a programme has caused certain effects by observing that the change after exposure to the programme is a noticeable departure from changes which were occurring anyway.

We may, however, still want to rely on control groups but to accept the fact that they have to be created in some non-random fashion. The *comparative change design* allows us to do just that. For example, all individuals who are eligible to receive programme benefits in a particular region or city may form the programme group, whilst individuals in another region or city become the control group. The key is to ensure that the two groups are assigned on the basis of factors which cannot reasonably be causes of the observed effects. However, there is always the possibility of selection bias. Very often, there are good reasons why some people participate in a programme whilst others who are also eligible do not. In the ERASMUS programme, beneficiaries tend to have better than average academic results and tend to come from families with high incomes and a high degree of international exposure. To compare ERASMUS students with students with lower grades or from families with a significantly different socio-economic profile would not be appropriate. These factors may well provide alternative explanations for the effects which would otherwise be attributable to the programme.

The *criterion-population design* is an improvement on the comparative change design since, as is seen in Figure 4.2. it does not require the existence of a distinct control group. In the comparative change design, the programme and control groups are two distinct groups drawn from a larger population. In the

criterion-population design, however, the larger population itself is identified and used as the basis for comparison. In this case, the possibility of selection bias is confined to just one group - the programme group. The evaluator need only worry that the programme group, without exposure to the programme, may not be representative of the larger population. This design is particularly appropriate where the evaluator cannot easily create a control group but does have access to information about the larger population from which the programme group is drawn.

4.1.4. The descriptive approach to evaluation design

The causal approach to evaluation design is appropriate to situations where the evaluator needs to arrive at a defensible, usually quantitative, estimate of the counterfactual situation in order to determine whether observed effects have indeed been caused by a programme. It is not, however, appropriate to all situations. Very often, the evaluator is instead concerned with providing a thorough description of a programme, including a descriptive study of its supposed effects. In this case, it is appropriate to choose a different type of evaluation design, one which is not rooted in the causal approach. Alternatively, an evaluator may find that the conditions necessary for adopting a causal evaluation design, which as Figure 4.2. illustrates are quite strong, simply do not exist. For example, many programmes have universal coverage, i.e. all members of the eligible population are programme beneficiaries (such as the Common Agricultural Policy, where all eligible farmers are beneficiaries). For such programmes, a design based on control groups would not be possible. In this case, the evaluator may decide that it is more appropriate to follow the descriptive approach. Descriptive evaluation designs can still yield useful information about a programme.

A frequently used descriptive evaluation design is the *ex post facto design* (not to be confused with ex post evaluation). This design is used in situations where the evaluator has only limited options in terms of making comparisons. He cannot decide which subjects are to be exposed to the programme and which are not, or what degree of exposure each is to receive. This is important for programmes which may have varying degrees of take-up, e.g. across regions. Furthermore, the evaluator can only refer to measurements of beneficiaries after their exposure to the programme, hence the use of the term “ex post facto”. In principle, it is still possible to arrive at an estimate of the counterfactual situation. If sample sizes are large enough, a statistical analysis could be performed to relate the various levels of programme exposure to the differences in observed effects, whilst controlling for other influences. A common problem, however, is that any relation which is identified may be spurious rather than real. Nevertheless, ex post facto designs have been used extensively to examine programmes which have been available in the past to the whole of the relevant population (programmes with universal coverage).

There is also a range of descriptive designs which may be called *case study designs*. Case studies are considered below as a data collection technique, which can be used in combination with other methods for obtaining data. Nevertheless, it is often the case that an evaluation design will be based on an in-depth study of one or more specific cases or situations. Case study designs are frequently used in situations where the programme being evaluated is highly

complex, where a comprehensive understanding is required on how the programme works or when an explanation is needed for a large pattern of heterogeneous effects.

Case study designs based on a single case may be appropriate where there is no requirement to generalise from the findings (i.e. where external validity is not a problem), or where we need to examine one critical instance or situation in detail. However, they are unlikely to be appropriate to situations where it is necessary to ask whether conclusions can be applied to a larger group. In this case, it is usual to rely on an evaluation design based on multiple cases. With such multiple case study designs, the key task facing the evaluator is to arrive at a defensible selection of cases to study, whilst ensuring some degree of variability among the cases so that they are representative.

4.2. Data collection techniques

> "Get your facts first, and then you can distort them as much as you please."
>
> Rudyard Kipling, *From Sea to Sea*

Relationships between a programme and its effects can only be established if data are available. Data can be defined as known facts which are used as a basis for inference. The most immediate source of data about a programme should normally be the monitoring system. However, the monitoring data will usually be restricted to outputs. In most cases, this will not be sufficient. Other data will have to be collected. The choice of technique for collecting data follows from the choice of evaluation design. In this section we will review some of the main data collection techniques which are used in programme evaluation. Before doing so, however, we will briefly describe different ways of classifying data.

The data collection techniques we will examine are *surveys, case studies, natural observations, expert opinion, reviews of programme documents* and *literature reviews*.

4.2.1. Classifying data

Data are said to be **subjective** where they involve personal feelings, attitudes and perceptions, and **objective** where they relate to observable facts that, in theory at least, do not concern personal opinions.

Data are described as **quantitative** when they involve numerical observations (e.g. the number of units of a specific good or service provided by the programme, the amount of the programme's budget spent on achieving a given objective, the number of beneficiaries of a programme, the rate of take-up of programme outputs). **Qualitative** data are non-numerical and related to categories (e.g. the gender of programme beneficiaries, their geographical location, etc.). Both subjective and objective data can be measured either qualitatively or quantitatively.

Collecting qualitative data on a programme (e.g. the opinion of experts, beneficiaries or programme administrators) is by no means inconsistent with the pursuit of analytical rigour mentioned at the beginning of this chapter. Indeed, besides the fact that many important aspects of programmes often do not lend themselves well to quantification, **qualitative data may be indispensable** to the correct interpretation of numerical information. Moreover, quantitative data which are supposedly 'objective' may turn out to be less than reliable e.g. if there are mistakes in measuring important variables (known as measurement error).

Another way of classifying data is to distinguish between **longitudinal** data, which are collected over time, and **cross-sectional** data, which are collected at the same point in time but from a variety of different geographical areas, etc.

A final classification is between **primary** data and **secondary** data. Primary data are taken directly from original sources or collected first hand. Secondary data, on the other hand, are data that have undergone extensive manipulation and interpretation.

The accuracy of data should be a key concern to those who are conducting an evaluation as well as to those who are sponsoring it. One should always be aware of the possibility of measurement errors. In addition, some definitions may not be entirely neutral.

Most evaluations will use a combination of data techniques both to address a wide range of issues and so that the weaknesses associated with one technique can be compensated for by the strengths of another. We will now examine each of these techniques in more detail.

4.2.2. Surveys

Surveys are used extensively in evaluations. They are a versatile way of collecting primary data, whether qualitative or quantitative, from a sample drawn from a wider population. A basic aim when conducting a survey is to aggregate and generalise results obtained from the sample to the wider population, so that conclusions can be drawn about units of the population which are not contained in the sample as well as elements that are.

In order to do this, surveys often rely on what is known as *probability sampling*, whereby each unit in the population has a known, nonzero probability of being selected for inclusion in the sample. The conclusions from this type of sample can then be projected, within statistical limits of error, to the wider population.

Survey information is usually acquired through *structured interviews* or *self-administered questionnaires*. The three main ways of obtaining data in a survey are by mail, telephone and face-to-face interviews. Since the evaluator needs to ensure that uniform data is collected from every unit in the sample, information tends to be collected in *close-ended form*, i.e. the respondent chooses from among pre-defined responses offered in the questionnaire or by the interviewer.

The two main types of survey are:

- ***cross-sectional surveys****- involve measurements made at a single point in time*. A cross-sectional survey may be the best approach when descriptive information is required for a large population. As well as being useful for acquiring factual information, cross-sectional surveys can also be employed to determine attitudes and opinions. On the other hand, it is difficult to use cross-sectional surveys when the information that is sought must be acquired by unstructured, probing questions and when a full understanding of events and conditions must be pieced together by asking different questions of different respondents.

- ***panel surveys****- involve measurements acquired at two or more points in time*. Panel surveys may be particularly appropriate where dynamic information (information about change) is required rather than static information. They can also be used for the purposes of causal inference, e.g. determining which of two related factors is the cause and which is the effect. On the other hand, panel surveys present their own administrative difficulties. The evaluator must be aware of the fact that the composition of the sample may change over time, and must avoid mistaking changes in the sample for changes in the conditions being assessed.

  Surveys can be a versatile method for collecting data. When they are properly done, they can produce reliable and valid information. Nevertheless, it should be pointed out that surveys have several drawbacks as a data collection technique. **They require expertise in their design, conduct and interpretation. If survey techniques are misused, the data obtained will be invalid and unreliable**.

  There is quite a large range of literature available on survey techniques and how to avoid the many pitfalls associated with the use of surveys, such as the various forms of bias and error which can occur.

4.2.3. Case studies

Case studies involve examining a limited number of specific cases or situations which the evaluator anticipates will be revealing about the programme as a whole. We have already discussed the use of case studies as an evaluation design. Here, we are concerned with the specific features of case studies as a data collection technique.

As a data collection technique, case studies tend to be appropriate where it is extremely difficult to choose a sample large enough to be statistically generalisable to the population as a whole; where generalisation is not important; where in-depth, usually descriptive data is required; and where the cases or projects to be studied are likely to be quite complex.

Instead of trying to obtain a statistically defensible sample (as with probability sampling when applied to surveys), the evaluator usually tries to obtain variety among the cases studied, in the hope that this will avoid bias in the picture that is constructed of the programme. A useful way of ensuring variety is to choose cases in function of a predetermined typology, which describes the main types of cases which need to be included.

The various stages involved in using case studies are:

- establish a typology of cases;
- selecting the cases and justifying the selection according to this typology;
- collecting all relevant information on each case;
- describing the cases, and highlighting important findings;
- comparing the various cases which have been chosen; and
- attempting to generalise from the chosen cases to other situations.

Case studies have the advantage of allowing the evaluator to pursue in-depth analysis but his sample will not be statistically defensible and it will therefore be difficult to generalise conclusions. Case studies can also be expensive and time-consuming to carry out. Finally, it should be pointed out that a researcher will usually not know whether or not a case study is representative until after he has conducted it.

4.2.4. Natural observations

This data collection technique involves the evaluator making on-site visits to locations where the programme is in operation and directly observing what is happening. Observational data can be used to describe the setting of the programme, the activities which take place in the setting, the individuals who participate in these activities (who may or may not be aware that they are being observed), and the meaning of these activities to the individuals.

The value of natural observations is that the evaluator can better understand programme activities and effects through observing first hand what is happening and how people are reacting to it. The evaluator will also have the chance to see things that may have escaped the programme administrators or things which they are reluctant to discuss in an interview. On the other hand, both the internal validity and the external validity (the generalisability) of the data may be limited since another person making the same on-site visit may derive different observations to those of the evaluator. In addition, there is the specific problem of the *Hawthorne effect*, which reminds us that programme staff and beneficiaries can behave quite differently from their normal patterns if they know that they are being observed (see Box 4.1. below).

Box 4.1. The Hawthorne Effect

In the late 1920s and early 1930s, research conducted at a factory in Hawthorne, Chicago in the United States found that output improved simply because experiments were taking place that persuaded workers that management cared about them. The term *Hawthorne effect* is used to explain situations where an experiment cannot be trusted because the very fact that the experiment is taking place is influencing the results obtained.

Scientists studying the effects of a new drug will often administer the drug to a treatment group and an inactive placebo to a control group. Neither group knows whether it has received the real drug or the placebo, so the potential Hawthorn effect should be eliminated. In practice, however, we can rarely be certain of this.

4.2.5. Expert opinion

Expert opinion relies on the necessarily subjective views of experts in a particular field as a source of data to address evaluation issues. Experts are chosen on the basis of their qualifications as well their knowledge and experience in a given area. There are various methods for systematising expert opinions, e.g. the Delphi technique, the Abacus of Régnier. For reasons of space, these are defined in the glossary contained in Annexe 1 of this guide.

Eliciting opinions from experts is really a specific type of survey, so the comments on surveys in section 4.2.2. above also apply here. However, as a data collection technique, expert opinion has certain specific strengths and weaknesses.

As far as strengths are concerned, expert opinion can be used to carry out measurements in areas where objective data are deficient. In addition, it tends to be a relatively inexpensive and quick data collection technique. On the other hand, like any subjective assessment, expert opinion presents a credibility problem. The evaluator may also experience difficulty in selecting a wide enough range or a large enough group of experts for use as a credible data source. Different stakeholders may dispute the claims of different experts. In any event, experts are unlikely to be equally knowledgeable about a particular area, so some sort of weighting system must be devised. Finally, the views of more outspoken experts may tend to stand out although their views may not be representative (this is referred to as “chatty bias”). For these reasons, the use of expert opinion as the sole data source should be avoided.

4.2.6. Reviews of programme documents

It will usually be possible for the evaluator to obtain information on the programme being evaluated by reviewing the general programme files, financial and administrative records and specific project documents. Any gaps in the available secondary data on file can then be identified and primary data collection methods used to complete the picture.

Programme document reviews can provide the evaluator with invaluable background information on the programme and its environment and hence put programme effects in context. It can produce a useful framework and basis for a subsequent primary data search. In addition, programme document reviews tend to be relatively quick and economical as a data collection technique. However, programme documents typically only shed light on programme outputs but not results or outcomes. More practically, they rarely yield information on control groups.

4.2.7. Literature reviews

Another source of secondary data is a literature review, which enables the evaluator to make the best use of previous work in the field under investigation and hence to learn from the experiences and findings of those who have carried out similar or related work in the past. There are two types of documents that can be used in a literature search. Firstly, there are published papers, reports and books prepared by academics, experts and official organisations. Secondly, there are specific studies in the area, including past evaluations.

A literature review is a relatively economical and efficient way of collecting secondary data. Furthermore, past research may suggest hypotheses to be tested, specific techniques for overcoming methodological difficulties or evaluation issues to be examined in the current study. The weaknesses of the literature review are those associated with the inherent nature of the secondary data. Data may not be relevant or compatible enough with the evaluation issues to be of use in the current study. Furthermore, the accuracy of secondary data is often difficult to determine. If a research synthesis has already been carried out as part of an evaluation project (see 3.2.5), then the evaluator should be made aware of this. Otherwise, there is a danger of replication.

4.3. Data analysis techniques

Evaluation is essentially an analytical activity. It involves analysing the data collected according to a given evaluation design and data collection technique in order to construct credible evidence about a programme. An understanding of the techniques used to analyse evaluation data is vital for drawing valid conclusions about programmes. This section provides a brief analysis of some of the main data analysis techniques that can be used in evaluations. Since some of the proposed methods are quite complex, it is not possible to provide anything more than a cursory summary of the different techniques and their strengths and weaknesses.

4.3.1. Statistical analysis

The use of statistics as a means of data analysis is very common in evaluation. Statistical analysis is used often used to describe phenomena in a concise and revealing manner. This is known as ***descriptive statistics***. It can also be used to test for relationships among variables or generalise findings to a wider population. This is known as ***statistical inference***.

Reporting the findings of an evaluation almost always involves the use of a certain amount of descriptive statistics. In addition to presenting and describing data in terms of tables and graphs, evaluators will frequently make use of common statistics such as the *mean* and the *standard deviation*.

The ***mean*** tells us the average of a set of values. For example, we may wish to know the average number of weeks before a long-term unemployed worker finds a new job after completing a training programme. The ***standard deviation*** is a measure of dispersion. Suppose we are interested in comparing two different training programmes aimed at two non-overlapping groups of long-term unemployed workers. With the first programme, many workers found a new job immediately after finishing the training, whilst many others found a new job only after more than a year had elapsed. With the second programme, most workers found a new job between four and eight months after finishing the training. The average time before a worker finds a new job may be the same for both programmes (i.e. they may have identical means), but it is clear that the standard deviation in the first programme is greater than in the second, because values are more dispersed around the mean.

There are a whole range of other statistics which can be used to describe data. Moving beyond descriptive statistics, evaluators also use statistical inference methods in order to establish relationships between variables, to estimate the

strength of any apparent relationship and to generalise conclusions to a wider population.

For example, suppose we wish to know whether the variation in the number of road accidents on any one day between two cities of a roughly equivalent size is due to chance or whether there are, in fact, systematic differences that need to be explained. A frequently used technique in statistics is *analysis of variance*, or ANOVA (for ANalysis Of VAriance), which is based on comparing the variance between samples with the variance within samples. To form our two samples, we would count the number of road accidents in each city on a selected number of days. This allows us to compare the variation in road accidents between cities with the variation in road accidents within cities.

Methods such as *regression analysis* can be used to establish the significance of any correlation (association) between variables of interest, e.g. the gender of a long-term unemployed worker and the amount of time before he or she finds a new job after a training programme. In regression analysis, we attempt to establish whether the variation in one variable (known as the *dependent variable*) can be explained in terms of the variation in one or more *independent variables*. The dependent variable is often quantitative, e.g. a person's income can be regressed on his educational qualifications, number of hours worked per week, age, etc. Special techniques are available, however, to deal with situations in which the dependent variable is qualitative, e.g. whether or not a person owns a car can be regressed on income, wealth, age, gender etc.

It should be noted that correlation does not imply causality. **Causality, in the commonly understood sense of the term, can never be proved statistically, although it may be very strongly suggested**. In the ANOVA example above, for example, we cannot prove that the difference in road accidents between the two cities is due to the fact that only one city benefited from the road safety campaign. It is up to the evaluator to present convincing arguments with which to discount the plausible alternatives (i.e. the threats to internal validity) to the programme as causes for the observed effects.

The strengths of statistical analysis as a data analysis technique are that it offers a valid way of assessing the statistical confidence the evaluator has in drawing conclusions from the data, and allows the findings of an evaluation to be summarised in a clear, precise and reliable way. On the other hand, not all programme effects can be analysed in a statistical manner. Furthermore, good statistical analysis requires a certain degree of expertise. The way data are categorised can distort as well as reveal important differences. The users of statistical analysis must be aware of the assumptions as well as the limitations of the particular statistical technique employed, as well as any problems with the reliability or validity of the data which have been used.

4.3.2. The use of models

Taking the use of statistical methods one stage further, the evaluator may wish to develop an analytical model in order to represent how a programme changes important socio-economic variables. Such models are normally taken from previous research. The main types of models are:

- input-output models - allow a researcher to examine systematically the linkages between the different parts of an economy, as the inputs of one industry can be thought of as the outputs of other industries.
- microeconomic models - are designed to examine the behaviour of households and firms in specific industries and markets, using equations which represent the supply and demand functions for a particular good or service.
- macroeconomic models - are used to model the behaviour of the economy as a whole and the evolution of important macroeconomic variables (such as inflation, employment, growth and the trade balance) over time.
- statistical models - are frequently used to examine relationships between specific programme effects. They are more versatile than the other types of model but are often less generalisable.

The main point to bear in mind about the use of models in evaluation is that it is important to determine the assumptions upon which the model is based, in order to understand and interpret correctly the information derived from it. Models are simplified representations of the real world. Simplification is necessary in order to isolate and focus on the effects of a programme. However, simplification can also lead to misinterpretation. The evaluator must exercise sound judgement if a model is to be correctly used.

A particular problem with macroeconomic models is their lack of *robustness*. In other words, a small change in the assumptions underlying the model can lead it to generate very different results. To get round this problem, a sensitivity analysis is normally conducted. Otherwise, several different models can be used to see whether their results converge.

4.3.3. Non-statistical analysis

Non-statistical analysis is carried out, for the most part, on qualitative data and is typically used in conjunction with statistical analysis of quantitative data. The use of non-statistical analysis should include an assessment of the *reliability* of any findings derived from such methods. In addition, the evaluator must exercise professional judgement to assess the *relevance* and *significance* of the available data to the evaluation issues at hand.

The major advantages of non-statistical data analysis are that many hard-to-quantify issues and concepts can be addressed and a more global viewpoint arrived at, often in a relatively inexpensive fashion. The major disadvantage is that conclusions based on non-statistical analysis will depend on the credibility of the evaluator and the logic of the arguments he presents. In any event, conclusions based solely on non-statistical analysis will not be as credible as conclusions derived from multiple lines of evidence and analysis.

4.3.4. Judgement techniques

Lastly, we will consider three specific analytical techniques which can be used to form judgements about programmes. Their use is more frequent in the ex ante evaluation of programmes, but they are often a useful way of forming judgements in intermediate or ex post evaluations. The three techniques are cost-benefit analysis, cost-effectiveness analysis and multi-criteria analysis.

*In **cost-benefit analysis,** a researcher compares all social and private costs and benefits of a programme with a view to determining whether the benefits exceed the costs, and if so by how much.* A key difficulty encountered in this approach is in the valuation of social costs and benefits. Social costs (such as the loss of an area of outstanding natural beauty) and social benefits (such as a reduction in road traffic accidents) usually have to be measured by some indirect means and converted into monetary values so that a comparison can be made with private costs and benefits.

Furthermore, it may not be appropriate to use prevailing market prices. Consider a situation of very high unemployment. In this case, the real cost of labour may be much lower than the prevailing market wage. The *opportunity cost* (the next best use of the otherwise unemployed workers had the project not gone ahead - some may have found jobs anyway, but many would have remained unemployed) is lower than the prevailing wage rate, and this low opportunity cost has to be represented by a *shadow price* which has to be derived somehow. Finally, once the monetary values for all private and social costs and benefits have been established, they have to be discounted to a common point in time. The appropriate interest rate which can be used to discount the various costs and benefits has to be chosen very carefully.

*In **cost-effectiveness analysis**, the researcher seeks to quantify the costs and benefits associated with a programme on the basis of the same principles which apply to cost-benefit analysis, but there is no requirement to transfer benefits into common monetary units.* A cost-effectiveness analysis of the road safety awareness programme discussed previously might discover that each 1,000 ECU of programme expenditure results in an average reduction of *X* road accidents per year. Thus, unlike in cost-benefit analysis, there is no requirement to convert the benefit (the reduction in road accidents) into monetary units.

Whether or not a programme is cost-effective depends on whether it outperforms other competing programme in reaching given objectives for less cost. For example, if the objective is to reduce traffic accidents in a given local area by a certain amount, the level of costs associated with meeting this objective through the road safety awareness programme could be compared with the costs of meeting it through lowering the speed limit or by introducing more traffic lights, pedestrianised areas and speed bumps. Thus, cost-effectiveness analysis is a particularly useful technique where a comparison between alternative ways of meeting the same objectives is called for.

In addition to the methodological problems which we have already discussed, it must be noted that neither cost-benefit analysis nor cost-effectiveness analysis can be used to *explain* particular results and outcomes. Nor do they have much to say about the distributional effects of a programme, i.e. who gains and who loses, and by how much.

Somewhat distinct from these two methods is ***multi-criteria analysis**,* which is essentially a decision-making tool which can be adapted to form judgements about programmes. Multi-criteria analysis allows us to formulate judgement on the basis of multiple criteria, which may not have a common scaling and which may differ in relative importance. Let us consider each of these elements in turn.

Programmes will normally have several different effects. If we are to form a judgement of a programme this means taking into account these multiple effects (e.g. the degree to which each of the specific objectives of the programme have been met). A problem is how to combine estimates of these effects when they do not have a common scaling, e.g. in the case of a Structural Funds programme we are typically interested in effects on employment (number of jobs created, number of jobs saved), on the enterprise base (number of new SMEs created), on the environment, and so on. How does one manage to combine all of these elements together in order to form a judgement on the programme as a whole? A further problem is that some criteria may be more important than others.

The technique of multi-criteria analysis allows key decision-makers to assign scores to the various criteria for judging a programme, which can then be weighted and used to compile an overall assessment of a programme.

Multi-criteria analysis has been used in the EU context in the case of the Structural Funds, but may not be easily transferable to other evaluation situations. It is, nevertheless, a useful technique.

Where to look for more information

The literature on evaluation designs is quite large, but two useful texts are Mohr (1995) and Treasury Board of Canada (1991). The latter also contains a good discussion of the various data collection and analysis techniques described here as well as an excellent bibliography. The standard reference on the use of case studies is Yin (1994). It is not possible to give a full list of introductory statistical texts in the space available here. A useful starting point is to consult the bibliography sections of evaluation texts. MEANS Handbook Number 4 on *Applying the Multi-criteria method to the Evaluation of Structural Programmes* is a useful introduction to this method in the specific context of the Structural Funds.

5. Reporting and disseminating evaluations

As we saw in Chapter 1, evaluations differ from ordinary research studies in that they are designed to be operationally **useful**. The usefulness of an evaluation will depend on its findings, conclusions and recommendations, and on how well these are reported and disseminated.

Reporting takes place when the evaluator transmits the evaluation (usually in the form of a document of activities and results) to the sponsors and when they, in turn, transmit a copy (or a summary thereof) to other interested parties within the Commission, including other services. *Dissemination* refers to the set of activities by which knowledge about an evaluation is made available to the world at large. This chapter examines how reporting and disseminating an evaluation can contribute to its utilisation.

Although reporting and disseminating have been left to the final chapter of this guide, **the sponsors of the evaluation need to start thinking about strategies for communicating the results of the evaluation at the same time as they are planning the evaluation itself**.

5.1. Maximizing the use of evaluations

In this section, we discuss some practical ways of ensuring the maximum use of evaluations. The first requirement is to **target the message to the audience**. This may seem obvious, but it is often overlooked when it comes to presenting and disseminating evaluations. So, when thinking about maximising the potential use of evaluations, it is important to be clear about the information needs of potential users of the evaluation.

These information needs will tend to differ according to whether the evaluation was conducted:

- to improve management;
- for reasons of accountability; or
- to assist in the allocation of budgetary resources.

An evaluation report which is primarily intended to *improve programme management* should be designed with a specialist audience in mind. For example, it can afford to be somewhat shorter and have a higher technical content than most evaluation reports. However, it may also be necessary to have a non-technical summary, perhaps written in a more discursive style, available for evaluation users who are not directly involved in the management of the programme, and who may lack specialist knowledge.

Evaluations conducted for reasons of *accountability* or to *assist in the allocation of budgetary resources* will normally have a more diverse range of potential users. For example, key decision-makers may have neither the time nor the

inclination to read complex pieces of analysis. A range of documents which present the same findings in different styles may be required. In any event, it is essential to have a self-contained *executive summary* available which can serve the information needs of senior Commission officials, Commissioners, Council representatives, Members of the European Parliament, the media, etc.

A second need is to **ensure that evaluation reports are timely**. In other words, the sponsors of the evaluation should ensure that reports are produced when they are most likely to be useful (e.g. in time to contribute to a decision on whether or not to renew a programme). This involves planning backward in time and making realistic projections of what must be done to meet any deadlines. In order to assist Commission services in this task, the Commission Communication on Evaluation adopted on 8 May 1996 introduced a requirement for all operational services to introduce their own *rolling programme* of evaluation work. Planned evaluations over a two-year time horizon are to be described in these rolling programmes and there should also be information on the decisions to which the evaluations are intended to contribute.

Finally, one should seek to **involve stakeholders in the design of the evaluation.** The evaluator and the sponsors can increase the potential usefulness of an evaluation by ensuring wide participation in the evaluation design. The aim is not only to ensure sensitivity to the interests of different stakeholders, but also to make them aware of future plans for utilising and disseminating the evaluation. This follows from the idea of evaluation as an *inclusive process*, as discussed in Chapter 3.

5.2. The presentation of the evaluation report

The evaluation report is the end product of the evaluation itself. It is important that it is well presented and well written.

5.2.1. The structure of the evaluation report

The evaluation report should follow a logical structure. Often, the precise structure (and sometimes the length as well) of the expected report will be specified in advance in the terms of reference. Box 5.1. below presents a typical structure for an evaluation report.

It is important to remember that there is no universally applicable evaluation report structure (although many DGs and services do have their own preferred structure for reports). Instead **what is important is that the structure of the report meets the needs of the sponsors of the evaluation as well as the principal stakeholders**. In the case of large programmes for which the evaluation task is to be divided among a number of external evaluators (e.g. broken down by region or country), it is obviously helpful for the different evaluation reports to have a common structure to facilitate reading and the preparation of any overall synthesis report.

Whilst we have stated that there is no universally applicable evaluation report structure, it is nevertheless important that all reports contain an **executive**

summary of no more than five pages in length. Ideally, this should feature towards the beginning of the report. **It should also be possible to circulate the executive summary as a separate stand-alone** document. It is the responsibility of the evaluation unit (or official in charge of evaluation) in each DG or service to ensure that a copy of the executive summary of all evaluation reports is sent to DG XIX. It is also useful for evaluation reports to contain a copy of their **terms of reference**.

Box 5.1. An example of an evaluation report structure

Title page:
- title and nature of evaluation (e.g. ex post)
- title of programme, generation, duration
- identification of author, date of submission, commissioning service

Table of contents:
- main headings and sub-headings
- index of tables of figures and graphs

Executive summary:
- an overview of the entire report in no more than five pages.
- a discussion of the strengths and weakness of the chosen evaluation design

Introduction:
- description of the programme in terms of needs, objectives, delivery systems etc.
- the context in which the programme operates
- purpose of the evaluation in terms of scope and main evaluation questions.
- description of other similar studies which have been done

Research methodology:
- design of research
- implementation of research and collection of data
- analysis of data

Evaluation results:
- findings
- conclusions
- recommendations

Annexes:
- terms of reference
- additional tables
- references and sources
- glossary of terms

5.2.2. The clarity of the evaluation report

In order for an evaluation to be useful, it must be understood. This is obviously the primary responsibility of the evaluator, but the sponsor may be called upon to defend the report to stakeholders and other audiences, and so the responsibility is to some extent shared.

A potential reader of an evaluation report must be able to understand:

- the purpose of the evaluation;

- exactly what was evaluated;
- how the evaluation was designed and conducted;
- what evidence was found;
- what conclusions were drawn; and
- what recommendations, if any, were made.

Writing an evaluation report can be challenging, as it calls for a variety of writing styles corresponding to the different parts of the report: a methodological part, descriptions of the programme and its effects, conclusions drawn from previous studies, analysis based on new findings and ensuing conclusions and recommendations.

On the one hand, the report must provide sufficient information in an analytically rigorous way to constitute a firm foundation for conclusions and recommendations. On the other hand, the report must be comprehensible to the intelligent non-specialist. This means keeping technical language to a minimum and explaining technical or unfamiliar concepts. A glossary of terms and other technical annexes can be useful in this respect.

It is likely that only a small proportion of the target audience will read the full report. It is therefore essential that the executive summary is well written. A frequent problem is that executive summaries are hastily prepared and so only give the reader a poor idea of the arguments and analysis contained in the main report. In other words, they are neither true “summaries”, nor do they permit “executive” decisions to be made.

Below is a list of other problems which can detract from the clarity of an evaluation report:

- failing to describe the programme being evaluated in sufficient detail (i.e. assuming that everyone who reads the evaluation report will be sufficiently acquainted with the programme and its rationale);
- failing to describe the methods used in the evaluation for the collection and analysis of data, to justify the choice of methods used or to indicate the strengths and weaknesses of the selected methods;
- using information without giving the source;
- arriving at findings which are not based firmly on evidence;
- reaching conclusions which are not explicitly justified (i.e. not systematically supported by findings), so that an independent reader cannot assess their validity; and
- making recommendations which are not adequately derived from conclusions.

5.3. The dissemination of evaluations

Dissemination encompasses the whole range of activities by which the information contained in evaluation reports is made available to wider audiences. Below is a list of stakeholder groups which can be potential audiences for an evaluation:

- **key policy-makers and decision-makers** - in the case of evaluations of EU programmes, this group can include the Commission, the European Parliament, the Council and national administrations;

- **programme sponsors** - normally, the relevant unit within the managing Directorate-General or service which is responsible for initiating and funding the programme to be evaluated;

- **evaluation sponsors** - organisations that initiate and fund the evaluation. (N.B. This group can be identical to the programme sponsors, depending on the specific features of the managing Directorate-General or service);

- **programme beneficiaries** - persons or groups who receive the goods and services provided by the programme being evaluated;

- **programme management** - persons and groups responsible for overseeing and co-ordinating the programme itself. In the case of many EU programmes, where day-to-day managerial tasks are contracted out to private entities, the programme management is often divorced from the programme sponsors; and

- **other interest groups and the academic community** - organisations, groups or individuals in the immediate environment of the programme, or having a general interest in the programme and its evaluation (e.g. the World-wide Fund for Nature in the case of many environmental programmes), and academics with a general scientific interest in the programme being evaluated.

Given the wide divergence between the potential audiences, it is obviously important that evaluation findings are communicated in ways that are appropriate to each one. Aside from circulating the full report, communication can take place through the circulation of the executive summary or through oral presentations based on audio-visual material.

When evaluators or sponsors wish to ensure dissemination of the information derived from an evaluation other than through distributing the report itself, their most important task is to **target the presentation to match the audience**. Box 5.2. below provides some of the main questions to ask when analysing the target audience of a presentation.

Box 5.2. Analysing the target audience

- ***How is the target audience composed?***
- ***What exactly do they need to know and why?***
- ***What is their knowledge of the evaluation?***
- ***Were they involved in the evaluation design? If so, to what extent? If not, why?***
- ***How might they be encouraged to attend any presentation?***
- ***What advantages and disadvantages might result to them from the evaluation?***
- ***Which evaluation questions are of interest to them?***
- ***What other issues are important to them?***
- ***Are they likely to object to particular findings, conclusions or recommendations?***
- ***How might these objections be overcome?***
- ***How interested will they be in the fine details compared to the overall picture?***

It is important to bear in mind that **different stakeholders are likely to react in different ways to a presentation of evaluation findings**.

Programme beneficiaries present particular problems. They are often unorganised and geographically scattered compared to other stakeholders. In the case of some programmes, beneficiaries may even be unwilling to identify themselves. Where they do make their voices heard, it may be through organisations which purport to represent their interests.

Finally, it is important to remember that **conflicts of interest are, to some extent, inevitable where there is a multiplicity of stakeholders**. The following points should therefore be borne in mind:

- conflicts of interest can best be tackled at the outset by having an inclusive management structure.
- by clearly separating findings, conclusions and recommendations, the evaluator can draw a line between the evidence that was found about a programme and his own personal opinions. Thus, even if some stakeholders choose to reject certain recommendations, they may be less inclined to dispute findings and conclusions.
- programme managers can, if need be, formulate their own observations on reports prepared by external experts.
- by no means should evaluation become entangled in negotiation. The professional expertise and conscience of an external evaluator should be a sufficient guarantee for the impartiality and credibility of his findings and conclusions.

Where to look for more information

A useful first source of information on strategies for reporting and disseminating evaluations will normally be the unit or official responsible for evaluation within each Directorate-General or service. The Joint Committee on Standards for Educational Evaluation (1994) contains a list of professional standards which evaluators should aim to meet. Many of these standards are applicable to reporting and disseminating evaluations. Although written with the evaluation of educational programmes in mind, the standards suggested in this text have a much wider potentially applicability. See also MEANS Handbook Number 1 on *Organising Intermediate Evaluation in the context of Partnerships*. Another helpful text is Rossi and Freeman (1993). Breakwell and Millward (1995) contains a very useful chapter on presenting evaluation findings.

References

[1] From Halcolm's *The Real Story of Paradise Lost* quoted in Patton (1986).

[2] HM Treasury (1988). UK government.

[3] Conseil scientifique de l'évaluation (1996).

[4] United Nations.

[5] Adapted from Rossi and Freeman (1993).

[6] MEANS Glossary.

[7] European Commission, Directorate-General for Development (1993).

[8] Viveret (1989).

[9] This diagram was adapted from one used by C3E, Lyon.

Annexe 1. Glossary of evaluation terms

Abacus of Régnier
A method for systematizing the opinions expressed in a group (e.g. of experts). The group is brought together and confronted with a list of closed questions to which each member must respond in a non-verbal manner using a seven colour code (two shades of green for signifying agreement, two shades of red for signifying disagreement, orange for signifying hesitation, white for indicating that the individual does not possess the information necessary to reply to the question and black for situations where the individual rejects the terms of the question). See also *Delphi technique*, *expert opinion*.

analysis
See *data analysis*.

analysis of variance
A widely-used statistical inference technique, based on comparing the variance between samples with the variance within samples. This can tell us whether there is any systematic difference between samples that needs to be explained. See also *sample*, *statistical analysis*, *variance*.

analytical agenda
A logical structure imposed on the different questions to be asked in an evaluation. This serves to transform the general, often vague, questions which those requesting the evaluation have in mind into questions which are precise enough to be manageable by evaluation research methods. Once the analytical agenda has been drawn up, those responsible for commissioning the evaluation have to ask whether the intervention is indeed evaluable in terms of this analytical agenda. See also *evaluability*, *evaluability assessment*, *evaluation project*, *intervention logic*.

ANOVA
See *analysis of variance*.

appraisal
See *ex ante evaluation*.

audit
A control function, which is primarily concerned with verifying the legality and regularity of the implementation of resources in a programme. Audit has traditionally covered areas such as the verification of financial records (financial audit). See also *performance audit*, *evaluation*.

before-and-after design
An example of a quasi-experimental design in which one simply compares the relevant state of the world after the intervention with its state beforehand and attributes any difference to the effects of the intervention. A particular weakness of this design is the possibility that something else besides the intervention accounts for all or part of the observed difference over time. See also *control group*, *counterfactual situation*, *evaluation design*, *internal validity*, *intervention logic*, *quasi-experimental designs*, *programme group*.

benchmarks
Standards by which the performance of an intervention can be assessed in a non-arbitrary fashion. An obvious way of deriving benchmarks would be to examine the intervention's objectives as expressed by expected outputs, results and outcomes. Ideally, benchmarks should allow us to compare the performance of an intervention with that of other policy instruments in the same field of action or in a related one. See also *general objectives*, *indicator*, *intervention*, *objectives*, *operational objectives*, *outcomes*, *outputs*, *results*, *specific objectives*.

case studies
A data collection technique involving the examination of a limited number of specific cases or projects which the evaluator anticipates will be revealing about the programme as a whole. Case studies tend to be appropriate where it is extremely difficult to choose a sample large enough to be statistically generalisable to the population as a whole; where generalization is not important; where in-depth, usually descriptive data is required; and where the cases or projects to be studied are likely to be quite complex. See also *case study designs*, *data collection*.

case study designs
A class of evaluation designs in the descriptive rather than the causal approach. It is often the case that an evaluation design will be based on an in-depth study of one or more specific cases or situations. See also *case studies*, *evaluation design*.

chatty bias
A general problem which arises when the views of more outspoken individuals (e.g. experts) tend to stand out, although their views may not be representative. See also *expert opinion*.

collection
See *data collection*.

comparative change design
An example of a quasi-experimental design in which any known or recognisable difference between the programme and control groups is taken into account in the statistical analysis. The problems with this design are, firstly, that there may be some other factor which explains some or all of the variation in the intervention and in the observed effects, and, secondly, that there may be initial differences between the programme and control groups which have an influence on observed effects and which can therefore become confounded with the influence of the programme on these effects (selection bias). See also *control group*, *counterfactual situation*, *evaluation design*, *internal validity*, *intervention logic*, *quasi-experimental designs*, *programme group*, *selection bias*.

control group
A group of subjects which have not been exposed to an intervention. The control group should resemble the programme group (the subjects which have been exposed to the intervention), so that systematic differences between the two groups may be attributed to the effects of the intervention once other plausible

alternative hypotheses have been eliminated or discounted. See also *counterfactual situation*, *evaluation design*, *intervention logic*, *programme group*.

cost-benefit analysis
A judgemental technique in which a researcher compares all social and private costs and benefits of a programme with a view to determining whether the benefits exceed the costs, and if so by how much. Social costs and social benefits usually have to be measured by some indirect means and converted into monetary values so that a comparison can be made with private costs and benefits. Furthermore, it may not be appropriate to use prevailing market prices. Consider a situation of very high unemployment. In this case, the real cost of labour may be much lower than the prevailing market wage. The opportunity cost (the next best use of the otherwise unemployed workers had the project not gone ahead) is lower than the prevailing wage rate, and this low opportunity cost has to be represented by a shadow price which has to be derived somehow. See also *cost-effectiveness analysis*.

cost-effectiveness analysis
A judgmental technique in which the researcher quantifies the costs and benefits associated with a programme on the basis of the same principles which apply to cost-benefit analysis, but there is no requirement to transfer benefits into common monetary units. See also *cost-benefit analysis*, *effectiveness*.

counterfactual situation
The situation which would have arisen had the intervention not taken place. In order to derive the counterfactual situation we need an evaluation design. Except for the theoretical case of the ideal experimental design, we can never know the counterfactual situation with certainty. Real world evaluation designs tend to be based on an estimate of the counterfactual derived either from comparing subjects who were exposed to an intervention with a comparison group who were not exposed, or from examining subjects before and after exposure. See also *control group*, *evaluation design*, *ideal experimental design*, *intervention logic*, *programme group*.

criterion-population design
An example of a quasi-experimental design, which attempts to improve on the comparative change design. In the latter, the programme and control groups are two distinct groups drawn from a hypothetical larger population. In the criterion-population design, however, the hypothetical population is identified and used for the comparison group. In this case, the possibility of selection bias is confined to just one group - the programme group. This design is particularly appropriate where the evaluator cannot easily create a control group but does have access to information about the larger population from which the programme group is drawn. See also *control group*, *comparative change design*, *counterfactual situation*, *evaluation design*, *internal validity*, *intervention logic*, *quasi-experimental designs*, *programme group*, *selection bias*.

cross-sectional data
See *data*.

cross-sectional surveys
See surveys.

data
Known facts which can be used as a basis for inference. Subjective data involve personal feelings, attitudes and perceptions; objective data relate to observable facts. Quantitative data involve numerical observations; qualitative data are non-numerical and related to categories. Longitudinal data are collected over time; cross-sectional data are collected from the same point in time, but from a variety of different geographical areas, etc. Primary data are taken directly from original sources or collected first hand; secondary data have undergone extensive manipulation and interpretation. See also *data analysis*, *data collection*.

data analysis
The main techniques used to interpret information about an intervention for use in an evaluation are statistical analysis, the use of models, non-statistical analysis and judgement techniques, such as cost-benefit analysis, cost-effectiveness analysis and multi-criteria analysis. See also *cost-benefit analysis, cost-effectiveness analysis*, *data collection*, *models*, *multi-criteria analysis*, *non-statistical analysis*, *statistical analysis*.

data collection
The main techniques used to gather information about an intervention for use in an evaluation are surveys, case studies, natural observations, expert opinion, reviews of programme documents and literature reviews. See also *case studies*, *data analysis*, *evaluation design*, *expert opinion*, *literature reviews*, *natural observations*, *programme document reviews*, *surveys*.

deadweight
Deadweight is defined as effects which would have arisen even if the intervention had not taken place. Deadweight usually arises as a result of inadequate delivery mechanisms which fail to target the intervention's intended beneficiaries sufficiently well. As a result, other individuals and groups who are not included in the target population end up as recipients of benefits produced by the intervention. Deadweight is really a special case of programme inefficiency. See also *delivery mechanisms*, *efficiency*, *target population*.

delivery mechanisms
The organisational arrangements which provide the goods and services funded by the intervention to its intended beneficiaries, i.e. its target population. See also *target population*.

Delphi technique
A technique which can be used to systematise expert opinions. Experts are consulted separately in a number of different rounds. In each successive round, each individual is told the views of the other experts in the previous round. This technique can be used to arrive at a consensus, or at least to reduce disagreements. See also *Abacus of Régnier*, *expert opinion*.

dependent variable
See *regression analysis*.

descriptive statistics
See *statistical analysis*.

displacement
Displacement and substitution are two closely related terms which are used to describe situations where the effects of an intervention on a particular individual, group or area are only realised at the extent of other individuals, groups or areas. Consider, for example, the case of a programme to provide employment subsidies. In a firm which benefits from this programme, subsidised workers may take the place of unsubsidised workers who would otherwise have been employed by that firm. This is known as substitution. Alternatively, a firm benefiting from the employment subsidies may win business from other firms which do not participate in the scheme. Thus, the jobs created in the participating firm may be partly or wholly offset by job losses in other firms. This is known as displacement.

dissemination
The set of activities by which knowledge about an evaluation is made available to the world at large. See also *reporting*.

double-loop learning
A type of feedback, in which the information compiled by an evaluation is used to call into question the very existence of an intervention or to bring about major changes in its basic orientations. Double-loop learning is almost always the result of summative evaluations. It is of key importance in focusing the activities of the European Union towards meeting the evolving needs of its citizens. See also *feedback*, *formative evaluation*, *single-loop learning*, *summative evaluation*.

effectiveness
To what extent have the intervention's impacts contributed to achieving its specific and general objectives? See also *cost-effectiveness analysis*, *general objectives*, *impacts*, *intervention logic*, *objectives*, *outcomes*, *results*, *specific objectives*.

efficiency
How economically have an intervention's inputs been converted into outputs and results? See also *inputs*, *intervention logic*, *outputs*, *results*.

evaluability
The issue of whether or not the questions raised by a given analytical agenda for an evaluation are at all answerable by an evaluator using appropriate research methods. To know whether the questions can be answered with an acceptable degree of credibility, it is often advisable to perform an evaluability assessment. If an intervention is not evaluable in terms of this analytical agenda (e.g. because adequate data are not yet available), this can lead to a decision to postpone the evaluation or to draw up a new, more realistic analytical agenda. See also *analytical agenda*, *evaluability assessment*, *evaluation project*.

evaluability assessment
An attempt to determine whether or not the questions raised by a given analytical agenda for an evaluation are at all answerable by an evaluator using appropriate research methods. See also *analytical agenda*, *evaluability*, *evaluation project*.

evaluation
An in-depth study which takes place at a discrete point in time, and in which recognised research procedures are used in a systematic and analytically defensible fashion to form a judgement on the value of an intervention.

evaluation design
A model which is used to describe an intervention and provide evidence on the effects which may be attributable to it. Evaluation designs are either causal or descriptive in nature. A given design should lead to the choice of one or more data analysis and collection techniques. See also *counterfactual situation*, *data analysis*, *data collection*, *ideal experimental design*, *intervention logic*.

evaluation project
A sequence of logical steps starting out from the formulation of problems and interests motivating the evaluation to arrive at a series of questions that can be addressed in an analytically acceptable way. The aim is to establish a work plan setting out a framework in which the evaluation proper is to be conducted and then to choose the evaluator. There are seven steps involved in elaborating an evaluation project: (i) identifying the goals of the evaluation; (ii) delineating the scope of the evaluation; (iii) drawing up the analytical agenda; (iv) setting benchmarks; (v) taking stock of available information; (vi) mapping out the work plan; and, (vii) selecting the evaluator. See also *analytical agenda*, *benchmarks*, *management structure*, *research synthesis*, *scope*, *work plan*.

evaluation report
The end product of an evaluation, the evaluation report must follow a logical structure and meet the needs of the evaluation sponsors and the principal stakeholders. Evaluation reports must include an executive summary of no more than five pages in length. The structure of the expected report is usually specified by the sponsors in the terms of reference. See *dissemination*, *evaluation sponsors*, *executive summary*, *reporting*, *stakeholders*, *terms of reference*.

evaluation sponsors
The DG or service within the Commission responsible for launching the evaluation of an intervention. See also *management structure*, *organisational structure*, *stakeholders*, *steering group*, *terms of reference*.

ex ante evaluation
An evaluation conducted before the implementation of an intervention. Also referred to as an “appraisal”. See also *evaluation*, *ex post evaluation*, *intermediate evaluation.*

ex post evaluation
An evaluation conducted either on or after completion of an intervention. See also *evaluation*, *ex ante evaluation*, *intermediate evaluation.*

ex post facto design
An example of a descriptive design, which can be used where the evaluator cannot select who is to be exposed to the programme, and to what degree. These designs have been used to examine interventions with universal coverage. See also *control group*, *counterfactual situation*, *evaluation design*, *intervention logic*, *programme group*.

executive summary
It is likely that only a small proportion of the target audience will read the full evaluation report. It is therefore essential to produce a well-written executive summary of no more than five pages in length. This summary forms part of the report and can also be distributed as a stand-alone document. See also *evaluation report*.

experimental group
See *programme group*.

expert opinion
A data collection technique, similar to a survey, which relies on the necessarily subjective views of experts in a particular field. It is not recommended to rely on expert opinion as a sole data source, for example, because of problems with so-called “chatty bias”. See also *Abacus of Régnier*, *chatty bias*, *data collection*, *Delphi technique*, surveys.

external evaluation
An evaluation which is performed by persons outside the organisation responsible for the intervention itself. See also *evaluation*, *internal evaluation*.

external validity
The confidence one can have about whether or not one’s conclusions about the intervention can be generalised to fit circumstances, times, people, and so on, other than those of the intervention itself. A threat to external validity is an objection that the evaluation design does not allow causal inference about the intervention to be generalised to different times, places or subjects to those examined in the evaluation. See also *evaluation design*, *internal validity*, *intervention*, *intervention logic*.

feedback
The process by which the information compiled by an evaluation is used by decision-makers to either change the way in which an intervention is implemented, or to bring about a more fundamental change in the basic orientations of the intervention, including calling into question its very existence. See also *double-loop learning*, *single-loop learning*.

financial audit
See *audit*

formative evaluation
An evaluation concerned with examining ways of improving and enhancing the implementation and management of interventions. Formative evaluations tend to be conducted for the benefit of those managing the intervention with the intention of improving their work. See also *evaluation*, *summative evaluation*.

general objectives
The desired effects of an intervention expressed in terms of outcomes, i.e. the longer-term impact of the intervention on society (e.g. to reduce unemployment among the long-term unemployed). See also *intervention logic*, *objectives*, *operational objectives*, *outcomes*, *specific objectives*.

Hawthorne effect
The term "Hawthorne effect" is used to explain situations where an experiment cannot be trusted because the very fact that the experiment is taking place is influencing the results obtained. This reminds us that programme staff and beneficiaries can behave quite differently from their normal patterns if they know that they are being observed. See also *natural observations*.

ideal experimental design
A theoretical way of deriving the counterfactual situation, and hence the net impact of an intervention. It involves comparing two groups which are identical in all respects except one: exposure to the intervention. Differences between the group which has been exposed (the programme group) and the group which has not (the control group) are then attributable to the intervention. In the real world, this design does not exist since we can never be absolutely certain that the two groups are identical in all other respects. The potential non-equivalence of the two groups weakens the validity of any causal inference about the intervention. A number of real world evaluation designs are available which each have their own strengths and weaknesses. See also *control group*, *counterfactual situation*, *evaluation design*, *intervention logic*, *programme group*, *quasi-experimental designs*, *true experimental designs*.

impacts
A general term used to describe the effects of a programme on society. Impacts can be either positive or negative and foreseen or unforeseen. Initial impacts are called results, whilst longer-term impacts are called outcomes. See also *outcomes*, *results*.

independent variable
See *regression analysis*.

indicator
A characteristic or attribute which can be measured to assess an intervention in terms of its outputs or impacts. Output indicators are normally straightforward. Impact indicators may be more difficult to derive, and it is often appropriate to rely on indirect indicators as proxies. Indicators can be either quantitative or qualitative. The term "performance indicators" is also used. See also *benchmarks*, *general objectives*, *impacts*, *operational objectives*, *outputs*, *specific objectives*.

input-output models
See *models*.

inputs
The human and financial resources involved in the implementation of an intervention. See also *intervention*, *intervention logic*.

intermediate evaluation
An evaluation conducted during the implementation of an intervention. See also *evaluation*, *ex ante evaluation*, *ex post evaluation*.

internal evaluation
An evaluation which is performed by members of the organisation responsible for the intervention itself. See also *evaluation*, *external evaluation*.

internal validity
The confidence one can have in one's conclusions about what the intervention actually did accomplish. A threat to internal validity is an objection that the evaluation design allows the causal link between the intervention and the observed effects to remain uncertain. It may be thought of as a question of the following nature: could not something else besides the intervention account for the difference between the situation after the intervention and the counterfactual? See also *counterfactual situation*, *evaluation design*, *external validity*, *intervention*, *intervention logic*, *selection bias*.

interrupted time-series design
An example of a quasi-experimental design. It involves obtaining several measurements over time both before and after exposure to a programme in order to create a time series of observations. It is an improvement on the before-and-after design. See also *before-and-after design*, *control group*, *counterfactual situation*, *evaluation design*, *internal validity*, *intervention logic*, *quasi-experimental designs*, *programme group*.

intervention
A generic term used to cover all public actions. See also *policy*, *programme*, *project*.

intervention logic
The conceptual link from an intervention's inputs to the production of its outputs and, subsequently, to its impacts on society in terms of results and outcomes. The examination of the programme's intervention logic will be of central importance in most evaluations. The evaluator needs to ask how the programme achieves its specific objectives, and how do the specific objectives contribute to the attainment of the general objectives? The terms "theory of action", "programme logic" and "programme theory" are sometimes used to mean more or less the same thing. See also *general objectives*, *impacts*, *inputs*, *intervention*, *objectives*, *operational objectives*, *outcomes*, *outputs*, *results*, *specific objectives*.

interviews
See *surveys*.

literature reviews
A data collection technique which enables the evaluator to make the best use of previous work in the field under investigation and hence to learn from the experiences and findings of those who have carried out similar or related work in the past. There are two types of documents that can be used in a literature search. Firstly, there are published papers, reports and books prepared by academics, experts and official organizations. Secondly, there are specific studies in the area, including past evaluations. See also *data collection*, *research synthesis*.

longitudinal data
See *data*.

macroeconomic models
See *models.*

management structure
A hierarchical structure which allows for overall management of an evaluation, and, in particular, the evaluation project. As a minimum, such a management structure must involve the programme management (usually identical to the evaluation sponsors) and the unit, sector, or official inside the same DG in charge of evaluation. However, for an evaluation to be successful, it may be necessary to widen the management structure and create a steering group**.** See also *evaluation project*, *evaluation sponsors*, *organisational structure*, stakeholders, *steering group*.

mean
The most commonly used descriptive statistic, it tells us the average of a set of values. See also *standard deviation*, *statistical analysis*.

microeconomic models
See *models.*

models
There are various different models which seek to represent how an intervention changes important socio-economic variables. Such models are normally taken from previous research. The main types of models are: (i) input-output models, which allow a researcher to systematically examine the linkages between the different parts of an economy, as the inputs of one industry can be thought of as the outputs of other industries; (ii) microeconomic models, which are designed to examine the behaviour of households and firms in specific industries and markets using equations which represent the supply and demand functions for a particular good or service; (iii) macroeconomic models, which are used to model the behaviour of the economy as a whole and the evolution of important macroeconomic variables (such as inflation, employment, growth and the trade balance) over time; and, (iv) statistical models, which are frequently used to examine relationships between specific programme effects. See also *data analysis*, *statistical analysis*.

monitoring
The continuous process of examining the delivery of programme outputs to intended beneficiaries, which is carried out during the execution of a programme with the intention of immediately correcting any deviation from operational objectives. Evaluation, on the other hand, is carried out at a discrete point in time, and consists of an in-depth study. Monitoring often generates data which can be used in evaluations. See also *evaluation.*

multi-criteria analysis
A decision-making tool which can be adapted to form judgements about interventions. Multi-criteria analysis allows us to formulate judgements on the basis of multiple criteria, which may not have a common scaling and which may differ in relative importance.

natural observations
A data collection technique in which the evaluator makes on-site visits to locations where the intervention is in operation and directly observes what is happening. Observational data can be used to describe the setting of the intervention, the activities which take place in the setting, the individuals who participate in these activities (who may or may not be aware that they are being observed), and the meaning of these activities to the individuals. This form of data collection is particularly vulnerable to the Hawthorne effect. See also *data collection*, *Hawthorne effect*.

needs
The socio-economic problems which an intervention aims to address, expressed from the point of view of its target population. For example, the need to improve job opportunities for long-term unemployed workers who may suffer from a lack of relevant skills. See also *objectives*, *target population*.

non-statistical analysis
A general term used to describe the analysis of mainly qualitative data which is typically used in conjunction with statistical analysis (of either qualitative or quantitative data). Usually, this includes an assessment of the reliability of any findings derived from such methods. See also *data*, *data analysis*, *statistical analysis*.

objective data
See *data*.

objectives
The desired effects of an intervention. See also *general objectives*, *needs*, *operational objectives*, *specific objectives*.

operational objectives
The desired effects of an intervention expressed in terms of outputs, i.e. the goods and services produced by an intervention (e.g. to provide professional training courses to the long-term unemployed). See also *general objectives*, *intervention*, *intervention logic*, *objectives*, *outputs*, *specific objectives*.

opportunity cost
See *cost-benefit analysis*.

organisational structure
Specifying the evaluation's organisational structure, which is usually done in the terms of reference, involves delineating the role of different actors (especially important if the evaluation task is to be divided among different evaluators - for example, between internal and external evaluators), establishing reporting responsibilities (including, where appropriate, contact with evaluation steering groups, programme managers, other Commission services and Member State administrations) and identifying the procedure to be followed to disseminate and use the evaluation. See also *dissemination*, *evaluation project*, *external evaluation*, *feedback*, *internal evaluation*, *management structure*, *stakeholders*, *steering group*, *terms of reference.*

outcomes
The longer-term impact, usually expressed in terms of broad socio-economic consequences, which can be attributed to an intervention (e.g. a reduction in the number of long-term unemployed). See also *general objectives*, *impact*, *intervention*, *intervention logic*, *outputs*, *results*.

outputs
The goods and services produced by an intervention (e.g. training courses for the long-term unemployed). See also *intervention*, *intervention logic*, *operational objectives*.

panel surveys
See surveys.

performance audit
Conceptually closer to evaluation than traditional audit, performance audit is strongly concerned with questions of efficiency (of an intervention's direct outputs) and good management. Performance audit and evaluation share the same aim of improving the quality of programmes, but evaluation goes much further. It also looks at issues such as sustainability, relevance and the longer-term consequences of a programme. See also *audit*, *evaluation*.

performance indicator
See *indicator*.

policy
A set of activities, which may differ in type and have different direct beneficiaries, directed towards common general objectives. Policies are not delimited in terms of time schedule or budget. See also *general objectives*, *intervention*, *programme*, *project*.

population
In statistics, the entire aggregate of individuals or subjects, from which samples can be drawn. See also *sample*, *target population*.

primary data
See *data*.

probability sampling
A statistical technique used to obtain samples from a given population, whereby every unit in the population has a known, non-zero probability of being selected for inclusion in the sample. The conclusions from this type of sample can then be projected, within statistical limits of error, to the wider population. See also *population*, *sample*.

programme
A set of organized but often varied activities (a programme may encompass several different projects, measures and processes) directed towards the achievement of specific objectives. Programmes have a definite time schedule and budget. See also *intervention*, *project*, *policy*, *specific objectives*.

programme document reviews
A data collection technique based on reviewing general programme files, financial and administrative records and specific project documents. See also *data collection*.

programme group
A group of subjects which have been exposed to an intervention. The programme group can be compared with the control group (the subjects which have not been exposed to the intervention), in order to determine whether systematic differences between the two groups may be attributed to the effects of the intervention. See also *control group*, *counterfactual situation*, *evaluation design*, *ideal experimental design*, *internal validity*, *intervention*, *intervention logic*, *quasi-experimental designs*, *true experimental designs*.

programme logic
See *intervention logic*.

programme theory
See *intervention logic*.

project
A single, non-divisible public intervention directed towards the attainment of operational objectives, with a fixed time schedule and a dedicated budget. See also *intervention*, *programme*, *policy*, *operational objectives*.

qualitative data
See *data*.

quantitative data
See *data*.

quasi-experimental designs
A class of causal evaluation designs which take a more practical approach than is the case with true experimental designs. Control groups can still be used, but these have to be assigned through some non-random process. Alternatively,

one can examine beneficiaries before and after exposure to the intervention. See also *before-and-after design*, *comparative change design*, *control group*, *counterfactual situation*, *criterion-population design*, *evaluation design*, *ideal experimental design*, *interrupted time-series design*, *intervention logic*, *programme group*, *true experimental designs*.

questionnaires
See surveys.

randomised experimental designs
See *true experimental designs*.

Régnier's abacus
See *Abacus of Régnier*.

regression analysis
A statistical inference technique which can be used to establish the significance of any correlation (association) between variables of interest, e.g. the gender of a long-term unemployed worker and the amount of time before he or she finds a new job after a training programme. In regression analysis, we attempt to establish whether the variation in one variable (known as the dependent variable) can be explained in terms of the variation in one or more independent variables. The dependent variable is often quantitative, e.g. a person's income can be regressed on his educational qualifications, number of hours worked per week, age, etc. Special techniques are available, however, to deal with situations in which the dependent variable is qualitative, e.g. whether or not a person owns a car can be regressed on income, wealth, age, gender etc. See also *statistical analysis*.

relevance
To what extent are the intervention's objectives pertinent in relation to the evolving needs and priorities at both national and EU level? See also *intervention*, *intervention logic*, *needs*, *objectives*.

report
See *evaluation report*.

reporting
Reporting takes place when the evaluator transmits the evaluation report (usually in the form of a document, or else through some audio-visual presentation) to the sponsors and when they, in turn, transmit a copy (or a summary thereof) to other interested parties. See also *dissemination*, *evaluation report*, *evaluation sponsors*, *executive summary*.

research synthesis
An overview of the current state of knowledge about a socio-economic problem and about remedies through public policy, which is undertaken before an evaluation. This knowledge can be obtained from professional literature, media articles, administrative data, monitoring reports or published statistics. Preparing a research synthesis is often helpful prior to launching an evaluation. By listing the information that is available and comparing it with the needs ensuing from

the analytical agenda, the research synthesis will point to the principal information gaps which, in turn, set the data collection and analysis tasks to be undertaken by the evaluation. Reviews of literature can also be a data collection technique in the conduct of an evaluation. See also *analytical agenda*, *data analysis*, *data collection*, *evaluation project*, *literature reviews*.

results
The initial impact of an intervention (e.g. an improvement in the employability of the long-term unemployed through a rise in their skill level). See also *impact*, *intervention*, *intervention logic*, *outcomes*, *outputs*, *specific objectives*,

sample
A set of individuals or items selected from a given population so that properties and parameters of the population may be estimated, or so that hypotheses about that population may be estimated. See also *population*, *probability sampling*.

scientific studies
Whereas scientists may undertake research in order to expand the sum of human knowledge and frequently confine themselves to one highly specialised discipline, evaluations are undertaken for more practical reasons. Evaluations aim to inform decisions, clarify options, reduce uncertainties and generally provide information about programmes within their own specific contexts. See also *evaluation.*

scope
The field of investigation of an evaluation. Typically, this has to be defined from an institutional (EU versus national or local level), temporal (period review) and geographical (part of the EU territory) point of view. In addition, one has to identify the key evaluation issues (relevance, efficiency, effectiveness, utility, sustainability) which will be examined. See also *effectiveness*, *efficiency*, *evaluation project*, *relevance*, *sustainability*, *utility*.

secondary data
See *data*.

selection bias
Could not the differences between the control group and the programme group be due to initial differences in their characteristics rather than the effects of the intervention we are trying to evaluate? See also *control group*, *counterfactual situation*, *evaluation design*, *internal validity*, *programme group*.

shadow price
See *cost-benefit analysis*.

single-loop learning
A type of feedback, in which the information compiled by an evaluation is used to bring about changes in the way an intervention is implemented. Although single-loop learning is more often associated with formative evaluations, it can also arise in the case of summative evaluations. See also *double-loop learning*, *feedback*, *formative evaluation*, *summative evaluation*.

specific objectives
The desired effects of an intervention expressed in terms of results, i.e. the initial impact of the intervention on society (e.g. to improve the employability of the long-term unemployed by raising their skill level). See also *general objectives*, *intervention*, *intervention logic*, *objectives*, *operational objectives*, *organisational structure, results*, *specific objectives*.

sponsors
See *evaluation sponsors*.

stakeholders
The various individuals and organisations who are directly and indirectly affected by the implementation and results of a given intervention, and who are likely to have an interest in its evaluation (e.g. programme managers, policy-makers, the programme's target population). See also *evaluation sponsors*, *steering group*, *target population*.

standard deviation
A commonly used descriptive statistic, it provides a measure of dispersion for a set of values. See also *mean*, *statistical analysis*, *variance*.

statistical analysis
A commonly used data analysis technique. Statistical analysis is used often used to describe phenomena in a concise and revealing manner. This is known as descriptive statistics. It can also be used to test for relationships among variables or generalise findings to a wider population. This is known as statistical inference. See also *data collection*, *non-statistical analysis*.

statistical inference
See *statistical analysis*.

statistical models
See *models.*

steering group
Part of the management structure for an evaluation, a steering group allows other services (and possibly other stakeholders from outside the Commission) to contribute to the development of the evaluation project. See also *evaluation project*, *management structure*, *stakeholders*.

subjective data
See *data*.

substitution
See *displacement*.

summative evaluation
An evaluation concerned with determining the essential effectiveness of programmes. Summative evaluations tend to be conducted for the benefit of external actors (groups who are not directly involved in the management of a

programme), for reasons of accountability or to assist in the allocation of budgetary resources. See also *evaluation*, *formative evaluation*.

surveys
A widely-used technique for collecting data from a sample drawn from a given population. Surveys are often based on probability sampling, and survey information is usually obtained through structured interviews or self-administered questionnaires. Cross-sectional surveys involve measurements made at a single point in time. Panel surveys involve measurements acquired at two or more points in time. See also *data collection*, *population*, *probability sampling*, *sample*.

sustainability
To what extent can the programme's positive impacts (as measured by its utility) be expected to last after the intervention has been terminated? See also *impacts*, *intervention logic*, *outcomes*, *results*, *utility*.

target population
The intended beneficiaries (individuals, households, groups, firms) of an intervention. An intervention may have more than one target population. This term should be distinguished from "population" in the statistical sense. See also *intervention*, *population*, *stakeholders*.

terms of reference
The terms of reference outline the work to be carried out by the evaluator, the questions to be dealt with and the time schedule. They allow the sponsors of the evaluation to define their requirements and allow the evaluator to understand clearly what is expected of the work to be undertaken (including, often, the structure of the expected evaluation report). Clearly defined terms of reference are vitally important where an evaluation is to be conducted by an external expert, and can also be of tremendous use when it is to be performed in-house. See also *evaluation project*, *evaluation report*, *evaluation sponsors*, *external evaluation*, *internal evaluation*, *organisational structure*, *work plan*.

thematic evaluation
An evaluation which focuses on one or more themes which are common to several different interventions (programmes or other activities), for example, effects on the environment or on small and medium-sized enterprises.

theory of action
See *intervention logic*.

threat to external validity
See *external validity*.

threat to internal validity
See *internal validity*.

true experimental designs
The best real world approximations to the ideal experimental design, in which the evaluator tries to ensure the initial equivalence of the programme and

control groups by creating them beforehand through random assignment. Although causal inference based on such designs is usually very strong, true experimental designs are difficult to administer and implement. Also referred to as “randomised experimental designs”. See also *control group*, *counterfactual situation*, *evaluation design*, *ideal experimental design*, *intervention logic*, *programme group*, *quasi-experimental designs*.

utility
How do the programme’s impacts compare with the needs of the target population? This issue is closely related to sustainability. See also *impacts*, *intervention logic*, *needs*, *outcomes*, *results*, *sustainability*, *target population*.

variance
A descriptive statistic which provides a measure of dispersion. It is obtained by squaring the standard deviation. See also *analysis of variance*, *standard deviation*, *statistical analysis*.

work plan
A schema which identifies the investigations that need to be carried out by the evaluation in the light of the chief questions raised by the analytical agenda and the information gaps which have been identified. These investigations should be described in sufficient detail to provide a provisional picture of the data collection and analysis tasks lying ahead, as well as of the methodologies to be employed. In order to keep them manageable, it often proves useful to divide the various tasks to be done into different stages and to set a corresponding time-table for the delivery of the different evaluation parts. The work plan is also the appropriate place for costing the evaluation and its components. See *analytical agenda*, *data analysis*, *data collection*, *evaluation project*.

Annexe 2. Judging the quality of evaluation reports

An evaluation report will usually be the subject of a critical examination by several parties (e.g. the sponsors themselves, the principal stakeholders, DG XIX in the case of evaluations intended to contribute to decisions on whether or not to renew programmes). This should be taken into account in the evaluation design, and it will also be helpful if the evaluator is aware of this fact at the outset.

Below is a list of questions which DG XIX officials will typically ask (according to an established checklist) when judging evaluation reports submitted by different DGs and services:

- **Is the report well presented?**
  Overall, is it well organised and clearly written?
  Are features such as the description of the programme and the explanation of the research methodology presented transparently in the report?

- **Is the scope of the report adequate?**
  Does the report address the entire programme under consideration?
  Are links with other programmes discussed?
  Are expected outputs, results and outcomes examined?
  Is the programme's intervention logic analysed?
  Are any unforeseen results and outcomes addressed?
  Is the sustainability of the benefits generated by the programme assessed?
  Does the report deal with the question of whether the programme will still be relevant in the future?
  Does the report examine the budgetary aspects of the programme being evaluated and its cost-effectiveness?

- **Is the methodology of the report appropriate?**
  Did the evaluation design allow information (on outputs, results and outcomes) to be obtained that can reasonably be attributed to the programme?
  Were indicators used appropriately (distinguishing between outputs, results and outcomes)?
  Were any weaknesses of the employed methodology pointed out?

- **Are the report's conclusions and recommendations credible?**
  Are findings based firmly on evidence?
  Are conclusions systematically supported by findings?
  Are recommendations adequately derived from conclusions?

Annexe 3. Some evaluation *do's and don'ts*

Preparing and managing evaluations

Establishing a management structure

DO	DON'T
• Establish a management structure, involving at least the programme management and the unit or official in charge of evaluation within the same DG or service	
• Consider widening the management structure to create a steering group, involving other Commission services and significant stakeholders	• Don't allow the steering group to become too large. It may lose its role as a management body and degenerate into a negotiation forum
• Remember the need for the active participation of the management structure in the evaluation to deal with problems that may arise once it is underway	

Elaborating the evaluation project

Identify the goals of the evaluation

DO	DON'T
• Clearly specify why the evaluation is being conducted and who are its principal users	• Don't launch evaluations with goals which are not realistically attainable

Delineate the scope of the evaluation

DO
• Delineate the scope of the evaluation, i.e. define its field of investigation (from an institutional, temporal and geographical point of view) and identify which key issues (relevance, efficiency, effectiveness, utility and sustainability) are to be examined

Draw up the analytical agenda

DO	DON'T
• Formulate the analytical agenda by imposing a logical structure on the questions to be asked in the evaluation	• Don't forget to attempt to retrieve the programme's intervention logic, paying special attention to the main assumptions embedded in it
• Where the programme's general and specific objectives have to be reconstructed from scratch, this should be done transparently by the management structure, preferably under the responsibility of a steering group	
• Use the main stakeholders' impressions about the programme as "working hypotheses" to be critically examined	
• Consider whether the programme is evaluable in terms of the chosen analytical agenda (where necessary,	• Don't launch evaluations which definitely cannot be evaluated in terms of the chosen analytical agenda. However, even

perform an evaluability assessment)

if a programme is only partially evaluable, it may still be useful to proceed with the evaluation

Set benchmarks

DO	DON'T
• Try to identify some benchmarks against which the programme can be assessed	• Don't interpret data on benchmarks in a simplistic fashion: if a programme falls short of achieving its objectives, it may still be successful compared to other programmes or remedies which have been tried in the past

Take stock of available information

DO	DON'T
• Take stock of available information (e.g. by preparing a research synthesis). By comparing this with the needs arising from the analytical agenda, the main information gaps will be highlighted. This, in turn, sets the data collection and interpretation task to be carried out by the evaluation itself	• If it is foreseen that the evaluation will involve a literature review as a data collection technique, it may not be necessary to conduct a research synthesis

Map out the work plan

DO	DON'T
• Establish the tasks which need to be carried out by the evaluation in the light of the main questions raised by the analytical agenda and the information gaps which have been identified • Describe the tasks in sufficient detail • Where possible, divide the various tasks into different stages and set a corresponding time-table for the delivery of the various parts • Cost the evaluation and its components. For internal evaluations, estimate the time to be spent by officials and other administrative expenditure. For external evaluations, estimate costs before launching a call for tenders	• Don't make unrealistic demands on the evaluator. Otherwise, there is a risk of the evaluation arriving too late or not achieving what it set out to do

Select the evaluator

DO	DON'T
• Once it is clear what type of questions the evaluation needs to ask and what its budget and time schedule are, decide whether it should be performed internally or externally	• Don't rely on an evaluator's technical competence as the only selection criterion. Other important criteria include independence, the ability to work to deadlines and value-for-money

Drawing up the terms of reference

DO

- Define *clear* terms of reference for the evaluation. This is vitally important for external evaluations and can be of tremendous use for internal evaluations
- Terms of reference normally include:
    - the legal base and motivation for

the evaluation
- the uses and users of the evaluation
- the description of the programme to be evaluated
- the scope of the evaluation
- the main evaluation questions
- the methodologies to be followed in data collection and design
- the work plan, organisational structure and budget
- the expected structure of the final evaluation report

Conducting evaluations

Evaluation designs

DO	
• Choose an evaluation design on the basis of the main type of questions to be addressed by the evaluation • The choice of design should be explicitly justified, and any weaknesses associated with the chosen design should be identified • Remember that different designs can be combined if necessary • Try to involve stakeholders in the choice of design • Be aware of any threats to causal inference in the design chosen. Where possible, develop arguments and collect evidence about whether or not these threats are important	• Don't assume that only causal designs are valid. There are many situations where descriptive designs can be useful

Data collection

DO	DON'T
• Use proven techniques for collecting data and justify the choice of techniques on the basis of the problems posed by an evaluation • Always be concerned with the accuracy of data. There is always the possibility of measurement errors. In addition, some definitions may not be entirely neutral	• Don't rely on just one data collection technique. The advantage of using more than one technique is that the strengths of one can balance the weaknesses of another • A literature review may not be useful if a research synthesis has been already conducted

Data analysis

DO	DON'T
• Use proven techniques for analysing data and justify the choice of techniques on the basis of the problems posed by an evaluation	• Don't rely on just one data analysis technique. The advantage of using more than one technique is that the strengths of one can balance the weaknesses of

another

- Where models are used, determine the assumptions upon which they are based

Reporting and disseminating evaluations

Maximising the use of evaluations

DO

- Three suggestions for maximising the potential use of evaluations are:
  - try to target the message to the particular information needs of a given audience
  - ensure that reports are timely
  - where possible, involve stakeholders in the choice of evaluation design

The presentation of the evaluation report

The structure of an evaluation report

DO

- Ensure that the structure of the report meets the needs of the sponsors and the principal stakeholders
- Ensure that the report includes an executive summary. It should also be possible to circulate this as a separate document
- Ensure that the report includes a copy of the terms of reference

The clarity of the evaluation report

DO

- Ensure that a potential reader can understand:
  - the purpose of the evaluation
  - exactly what was evaluated
  - how the evaluation was designed and conducted
  - what evidence was found
  - what conclusions were drawn
  - what recommendations, if any, were made

DON'T

- Try to avoid the following problems which can detract from the clarity of a report:
  - executive summaries which are hastily written
  - describing the programme in insufficient detail
  - failing to describe the methods used for data collection and analysis
  - failing to justify the choice of methods or to indicate the strengths and weakness of the chosen design
  - using information without giving the source
  - arriving at findings which are not based firmly on evidence
  - reaching conclusions which are not systematically supported by findings
  - making recommendations which are not adequately derived from

conclusions

The dissemination of evaluations

DO	DON'T
• Communicate evaluation findings in ways which are appropriate to the information needs of the different stakeholders • Aside from circulating the full report, use the executive summary or other means e.g. oral presentations based on audio-visual material • Tackle potential conflicts of interest between stakeholders through an inclusive management structure • Ensure that findings, conclusions and recommendations are clearly separated • Where necessary, programme managers can formulate their own observations on reports prepared by external experts	• Don't allow evaluation to become entangled in negotiation

Select bibliography

Breakwell, Glynis M. and Lynne Millward (1995). *Basic evaluation methods. Analysing performance, practice and procedure*. Leicester: British Psychological Society.

Conseil scientifique de l'évaluation (1996). *Petit guide de l'évaluation des politiques publiques.* March. Paris: CSE.

European Commission (1993). *Project cycle management. Integrated approach and logical framework*. Directorate-General for Development.

European Commission (1995). *Common guide for monitoring an interim evaluation*. Structural Funds.

HM Treasury (1988). *Policy evaluation: a guide for managers*. London: Her Majesty's Stationary Office.

Joint Committee on Standards for Educational Evaluation (1994). *The programme evaluation standards.* Second edition. Thousand Oaks, CA: Sage.

MEANS (1995). *Auditing, monitoring and evaluation of European structural policies. Should they be separated or integrated?* October. Lyon: European Commission and C3E.

MEANS Handbook No. 1. *Organising intermediate evaluation in the context of partnerships.* Lyon: European Commission and C3E.

MEANS Handbook No. 4. *Applying the multi-criteria method to the evaluation of structural programmes*. Lyon: European Commission and C3E.

Mohr, Lawrence B. (1995). *Impact analysis for programme evaluation***.** Second edition. Thousand Oaks, CA: Sage.

Patton, Michael Quinn (1986). *Utilization-focused evaluation*. Second edition. Beverly Hills: CA: Sage.

Rossi, Peter H. and Howard E. Freeman (1993). *Evaluation. A systematic approach.* Fifth edition. Newbury Park, CA: Sage.

Treasury Board of Canada (1991). *Programme evaluation methods.*

Viveret, Patrick (1989). *L'évaluation des politiques et des actions publiques*, rapport au Premier ministre. Paris: La Documentation française.

Yin, Robert K. (1994). *Case study research. Design and methods*. Second edition. Newbury Park, CA: Sage.

Index

O

P

Q

R

S

T

U

W