Philosophies and types of evaluation research

Elliot Stern

In:

Descy, P.; Tessaring, M. (eds)

The foundations of evaluation and impact research

Third report on vocational training research in Europe: background report.
Luxembourg: Office for Official Publications of the European Communities, 2004
(Cedefop Reference series, 58)

Additional information on Cedefop's research reports can be found on:
http://www.trainingvillage.gr/etv/Projects_Networks/ResearchLab/

For your information:

- the **background report** to the third report on vocational training research in Europe contains original contributions from researchers. They are regrouped in three volumes published separately in English only. A list of contents is on the next page.
- A **synthesis report** based on these contributions and with additional research findings is being published in English, French and German.

  Bibliographical reference of the English version:
  Descy, P.; Tessaring, M. *Evaluation and impact of education and training: the value of learning*. Third report on vocational training research in Europe: synthesis report. Luxembourg: Office for Official Publications of the European Communities (Cedefop Reference series)
- In addition, an **executive summary** in all EU languages will be available.

The background and synthesis reports will be available from national EU sales offices or from Cedefop.

For further information contact:

Cedefop, PO Box 22427, GR-55102 Thessaloniki
Tel.: (30)2310 490 111
Fax: (30)2310 490 102
E-mail: info@cedefop.eu.int
Homepage: www.cedefop.eu.int
Interactive website: www.trainingvillage.gr

Contributions to the background report of the third research report

Impact of education and training

Preface

The impact of human capital on economic growth: a review
Rob A. Wilson, Geoff Briscoe

Empirical analysis of human capital development and economic growth in European regions
Hiro Izushi, Robert Huggins

Non-material benefits of education, training and skills at a macro level
Andy Green, John Preston, Lars-Erik Malmberg

Macroeconometric evaluation of active labour-market policy – a case study for Germany
Reinhard Hujer, Marco Caliendo, Christopher Zeiss

Active policies and measures: impact on integration and reintegration in the labour market and social life
Kenneth Walsh and David J. Parsons

The impact of human capital and human capital investments on company performance Evidence from literature and European survey results
Bo Hansson, Ulf Johanson, Karl-Heinz Leitner

The benefits of education, training and skills from an individual life-course perspective with a particular focus on life-course and biographical research
Maren Heise, Wolfgang Meyer

The foundations of evaluation and impact research

Preface

Philosophies and types of evaluation research
Elliot Stern

Developing standards to evaluate vocational education and training programmes
Wolfgang Beywl; Sandra Speer

Methods and limitations of evaluation and impact research
Reinhard Hujer, Marco Caliendo, Dubravko Radic

From project to policy evaluation in vocational education and training – possible concepts and tools. Evidence from countries in transition.
Evelyn Viertel, Søren P. Nielsen, David L. Parkes, Søren Poulsen

Look, listen and learn: an international evaluation of adult learning
Beatriz Pont and Patrick Werquin

Measurement and evaluation of competence
Gerald A. Straka

An overarching conceptual framework for assessing key competences. Lessons from an interdisciplinary and policy-oriented approach
Dominique Simone Rychen

Evaluation of systems and programmes

Preface

Evaluating the impact of reforms of vocational education and training: examples of practice
Mike Coles

Evaluating systems' reform in vocational education and training. Learning from Danish and Dutch cases
Loek Nieuwenhuis, Hanne Shapiro

Evaluation of EU and international programmes and initiatives promoting mobility – selected case studies
Wolfgang Hellwig, Uwe Lauterbach, Hermann-Günter Hesse, Sabine Fabriz

Consultancy for free? Evaluation practice in the European Union and central and eastern Europe Findings from selected EU programmes
Bernd Baumgartl, Olga Strietska-Ilina, Gerhard Schaumberger

Quasi-market reforms in employment and training services: first experiences and evaluation results
Ludo Struyven, Geert Steurs

Evaluation activities in the European Commission
Josep Molsosa

Philosophies and types of evaluation research

Elliot Stern

Abstract

This chapter considers different types of evaluation in vocational education and training (VET). It does so from two standpoints: debates among evaluation researchers and the way contexts of use and evaluation capacity shape evaluation in practice. The nature of VET as an evaluation object is discussed and theories of evaluation are located in wider debates about the nature of knowledge and philosophies of science. The various roles of evaluation in steering and regulating decentralised policy systems are discussed, as is the way evaluation itself is regulated through the development of standards and professional codes of behaviour.

Table of contents

List of tables

Tables

1. Introduction

Until quite recently, evaluation thinking has been centred in North America. Only over the last ten years have we seen the growth and spread of evaluation in Europe. This has been associated with a significant expansion of evaluations supported by the European Union (EU), especially in relation to Structural Funds (European Commission, 1999), and the establishment of evaluation societies at Member State and European levels. There has also been the beginnings of a tradition of evaluation publishing in Europe, including the emergence of a major new evaluation journal edited for the first time from a European base and with a high proportion of European content. Many factors account for the growth in evaluation activities in Europe in recent years (Leeuw et al., 1999; Rist et al., 2001; Toulemonde, 2001). These include both structural and management considerations. Expenditure pressures, both at national and European level, have increased demands for improved performance and greater effectiveness within the public sector. Furthermore, public action is becoming increasingly more complex both in terms of the goals of programmes and policies and the organisational arrangements through which they are delivered.

The decentralisation of public agencies, together with the introduction of results based management and other principles commonly described under the heading 'new public management', have created new demands for accountability; these are often in multi-agency and partnership environments. Without the bottom line of financial measures to judge success, new ways of demonstrating impacts and results are being demanded; so are new ways of regulating and steering decentralised systems. We see evaluation nowadays not only applied to programmes or policy instruments but also built into the routines of administration. This is often associated with standards that are set by policy-makers in relation to the performance of those expected to deliver public services. Standards, a concept central to evaluation, have also come to be applied to evaluation itself. Even evaluation is not free from the demands to deliver reliable, high quality output.

1.1. Scope of this chapter

It is against this background that this chapter on 'types and philosophies' of evaluation has been prepared. The main sections are as follows.

Chapter 2 begins by seeking to define, or, more accurately, review, attempts to define evaluation through the work of various scholars and experts. It allows us to clarify the main types of evaluation that are in use in different institutional and administrative settings, even though the writing of scholars and experts focus on 'pure types' when compared with evaluation as practised. This section also begins to highlight issues related to standards and their role in evaluation more broadly.

Chapter 3 then considers the nature of evaluation in the context of vocational education and training (VET) and addresses the question: what characterises evaluation in this domain; is it in any way distinctive? The section includes both specific consideration of VET and more general characteristics of evaluation objects and configurations.

Chapter 4 seeks to locate evaluation theory within the broader setting of the nature of theory in the philosophy of science. Many of the debates in evaluation are shaped by, and reflect, these wider debates.

Evaluation theory – narrowly conceived – is then reviewed in Chapter 5. In many ways theory within evaluation (as will be discussed) is a very particular construction, though this does not invalidate wider understandings of the role of theory.

The reality and practice of evaluation are then considered in Chapter 6, bringing together a substantial body of research into evaluation use and institutionalisation in order to understand better different types of evaluation *in situ*.

Evaluation standards and codes of behaviour and ethics for evaluators are reviewed in Chapter 7, drawing on experience in North America, Australasia and, more recently, Europe.

Finally, in Chapter 8, the discussion returns to the question of evaluation standards and the role they play both in the evaluation process and in governance and regulatory processes to steer institutions and promote policies and reforms.

2. Can evaluation be defined?

There are numerous definitions and types of evaluation. There are, for example, many definitions of evaluation put forward in handbooks, evaluation guidelines and administrative procedures, by bodies that commission and use evaluation. All of these definitions draw selectively on a wider debate as to the scope and focus of evaluation. A recent book identifies 22 foundation models for 21st century programme evaluation (Stufflebeam, 2000a), although the authors suggest that a smaller subset of nine are the strongest. Rather than begin with types and models, this chapter begins with an attempt to review and bring together the main ideas and orientations that underpin evaluation thinking.

Indicating potential problems with 'definition' by a question mark in the title of this section warns the reader not to expect straightforward or consistent statements. Evaluation has grown up through different historical periods in different policy environments, with inputs from many disciplines and methodologies, from diverse value positions and rooted in hard fought debates in philosophy of science and theories of knowledge. While there is some agreement, there is also persistent difference: evaluation is contested terrain. Most of these sources are from North America where evaluation has been established – as a discipline and practice – and debated for 30 or more years.

2.1. Assessing or explaining outcomes

Among the most frequently quoted definitions is that of Scriven who has produced an evaluation Thesaurus, his own extensive handbook of evaluation terminology: '"evaluation" refers to the process of determining the merit, worth or value of something, or the product of that process […] The evaluation process normally involves some identification of relevant standards or merit, worth or value; some investigation of the performance of evaluands on these standards; and some integration or synthesis of the results to achieve an overall evaluation or set of associated evaluations.' (Scriven, 1991; p. 139).

This definition prepares the way for what has been called 'the logic of evaluation' (Scriven, 1991; Fournier, 1995). This logic is expressed in a sequence of four stages:

(a) establishing evaluation criteria and related dimensions;
(b) constructing standards of performance in relation to these criteria and dimensions;
(c) measuring performance in practice;
(d) reaching a conclusion about the worth of the object in question.

This logic is not without its critics (e.g. Schwandt, 1997) especially among those of a naturalistic or constructivist turn who cast doubt on the claims of evaluators to know, to judge and ultimately to control. Other stakeholders, it is argued, have a role and this changed relationship with stakeholders is discussed further below.

The most popular textbook definition of evaluation can be found in Rossi et. al.'s book *Evaluation – a systematic approach*: 'Program evaluation is the use of social research procedures to systematically investigate the effectiveness of social intervention programs. More specifically, evaluation researchers (evaluators) use social research methods to study, appraise, and help improve social programmes in all their important aspects, including the diagnosis of the social problems they address, their conceptualization and design, their implementation and administration, their outcomes, and their efficiency.' (Rossi et al., 1999; p. 4).

Using words such as effectiveness rather than Scriven's favoured 'merit worth or value' begins to shift the perspective of this definition towards the explanation of outcomes and impacts. This is partly because Rossi and his colleagues identify helping improve social programmes as one of the purposes of evaluation. Once there is an intention to make programmes more effective, the need to explain how they work becomes more important. Yet, explanation is an important and intentionally absent element in Scriven's definitions of evaluation: 'By contrast with evaluation, which identifies the value of something, explanation involves answering a Why or How question about it or a call for some other type of understanding. Often, explanation involves identifying the cause of a phenomenon, rather than its effects (which is a major part of evaluation). When it is possible, without jeopardizing the main goals of an evaluation, a good evaluation design tries to uncover microexplanations (e.g. by identifying those components of the curriculum package that are producing the major part of the good or bad effects, and/or those that are having little effect). The first priority, however, is to resolve the evaluation issues (is the package any good at all, the best available? etc.). Too often the research orientation and training of evaluators leads them to do a poor job on evaluation because they became interested in explanation.' (Scriven, 1991, p. 158).

Scriven himself recognises that one pressure moving evaluation to pay greater attention to explanation is the emergence of programme theory, with its concern about how programmes operate so that they can be improved or better implemented. A parallel pressure comes from the uptake of impact assessment associated with the growth of performance management and other managerial reforms within public sector administrations. The intellectual basis for this work was most consistently elaborated by Wholey and colleagues. They start from the position that evaluation should be concerned with the efficiency and effectiveness of the way governments deliver public services. A core concept within this approach is what is called 'evaluability assessment' (Wholey, 1981). The starting point for this assessment is a critical review of the logic of programmes and the assumptions that underpin them. This work constitutes the foundation for most of the thinking about programme theory and logical frameworks. It also prefigures a later generation of evaluation thinking rooted more in policy analysis that is concerned with the institutionalisation of evaluation within public agencies (Boyle and Lemaire, 1999), as discussed further below.

These management reforms generally link interventions with outcomes. As Rossi et al. recognise, this takes us to the heart of broader debates in the social sciences about causality: 'The problem of establishing a program's impact is identical to the problem of establishing that the program is a cause of some specified effect. Hence, establishing impact essentially amounts to establishing causality.' (Rossi et al., 1999).

The difficulties of establishing perfect, rather than good enough, impact assessments are recognised by Rossi and colleagues. This takes us into the territory of experimentation and causal inference associated with some of the most influential founders of North American evaluations such as Campbell, with his interest in experimental and quasi-experimental designs, but also his interest in later years in the explanatory potential of qualitative evaluation methods. The debate about experimentation and causality in evaluation continues to be vigorously pursued in various guises. For example, in a recent authoritative text on experimentation and causal inference, (Shadish et al., 2002) the authors begin to take on board contemporary criticisms of experimental methods that have come from the philosophy of science and the social sciences more generally. In recent years, we have also seen a sustained realist critique on experimental methods led in Europe by Pawson and Tilley (1997). But, whatever their orientations to experimentation and causal inference, explanations remain at the heart of the concerns of an important constituency within evaluation.

2.2. Evaluation, change and values

Another important strand in evaluation thinking concerns the relationship between evaluation and action or change. One comparison is between 'summative' and 'formative' evaluation methods, terms also coined by Scriven. The former assesses or judges results and the latter seeks to influence or promote change. Various authors have contributed to an understanding of the role of evaluation and change. For example, Cronbach (1982, 1989) rooted in policy analysis and education, sees an important if limited role for evaluation in shaping policy 'at the margins' through 'piecemeal adaptations'. The role of evaluation in Cronbach's framework is to inform policies and programmes through the generation of knowledge that feeds into the 'policy shaping community' of experts, administrators and policy-makers. Stake (1996) on the other hand, with his notion of 'responsive evaluation', sees this as a 'service' to programme stakeholders and to participants. By working with those who are directly involved in a programme, Stake sees the evaluator as supporting their participation and possibilities for initiating change. This contrasts with Cronbach's position and even more strongly with that of Wholey (referred to earlier) given Stake's scepticism about the possibilities of change at the level of large scale national (or in the US context Federal and State) programmes and their management. Similarly, Patton, (1997 and earlier editions) who has tended to eschew work at programme and national level, shares with Stake a commitment to working with stakeholders and (local) users. His concern is for 'intended use by intended users'.

Virtually everyone in the field recognises the political and value basis of much evaluation activity, albeit in different ways. While Stake, Cronbach and Wholey may recognise the importance of values within evaluation, the values that they recognise are variously those of stakeholders, participants and programme managers. There is another strand within the general orientation towards evaluation and change which is decidedly normative. This category includes House, with his emphasis on evaluation for social justice and the emancipatory logic of Fetterman et al. (1996) and 'empowerment evaluation'. Within the view of Fetterman and his colleagues, evaluation itself is not undertaken by external experts but rather is a self-help activity in which – because people empower themselves – the role of any external input is to support self-help. So, one of the main differences among those evaluators who explicitly address issues of programme and societal change is in terms of the role of evaluators, be they experts who act, facilitators and advocates, or enablers of self help.

2.3. Quantitative and qualitative methods

Much of the literature that forms the foundation of the explanatory strand in evaluation is quantitative, even though we have noted that the later Campbell began to emphasise more the importance of qualitative understanding. However, among those concerned with formative, responsive and other change-oriented evaluations, there is a predominance of qualitative methods. The scope of 'qualitative' includes processes as well as phenomena, which by their nature require qualitative description. This would include, for example, the means through which a programme was being implemented as well as the dynamics that occur during the course of evaluation (e.g. learning to do things better, improving procedures, overcoming resistance).

Stake emphasises qualitative methods. He is often associated with the introduction of case studies into evaluation practice, although he also advocates a full range of observational, interview and conversational techniques (Stake, 1995). Patton's commitment to qualitative methods is reinforced by his interest in the use by programme managers of the results of evaluations. 'They must be interested in the stories, experiences and perceptions of programme participants' (Patton, 2002; p. 10).

2.4. Evaluation types

After this 'tour' around some of the main arguments and positions in evaluation, it becomes possible to return to the matter of definition and types of evaluation. This is not a simple or single definition but types of evaluation can be seen to cohere around two main axes. The first axis is methodological and the second concerns purposes.

In terms of methodologies, looking across the different approaches to evaluation discussed above, we can distinguish three methodological positions:

(a) the criteria or standards based position, which is concerned with judging success and performance by the application of standards;
(b) the causal inference position, which is concerned with explaining programme impacts and success;
(c) the formative or change oriented position, which seeks to bring about improvements both for programmes and for those who participate in them.

Alongside these methodological distinctions are a series of definitions that are concerned with evaluation purposes. Distinguishing evaluation in terms of purpose has been taken up by many authors including Vedung (1997), evaluators at the Tavistock Instititute (Stern, 1992; Stern et al., 1992) and Chelimsky (1995, 1997). Most of these authors distinguish between different evaluation purposes that are clearly consistent with the overview presented above. Along this axis, we can distinguish between the following purposes:

(a) accountability, where the intention is to give an account to sponsors and policy-makers of the achievements of a programme or policy;

(b) development, where the intention is to improve the delivery or management of a programme during its term;

(c) knowledge production, where the intention is to develop new knowledge and understanding;

(d) social improvement, where the intention is to improve the situation of the presumed beneficiaries of public interventions.

There is a degree of correlation between these two axes as Table 1 suggests.

Table 1: **Overlaps between methodology and purpose**

Purposes	**Methodology**		
	Criteria and standards	Causal inference	Change orientation
Accountability	Outcome and impact evaluations. Mainly summative		
Development			Formative evaluation of programmes
Knowledge production		'What works' – improving future policy/practice	
Social improvement			Empowerment and participative evaluations

Evaluation for the purpose of accountability tends to be concerned with criteria and standards (or indicator studies). Development evaluations use change oriented methods to pursue the desired improvements in programme delivery. Evaluations for the purpose of knowledge production are often concerned with drawing causal inference from evaluation data. Finally, evaluations for the purposes of social improvement are also preoccupied with change oriented methods, though to improve the circumstances of programme participants and citizens rather than programme management *per se*. However, this is not to suggest a one-to-one association between methodologies and purposes.

In the world of evaluation in practice, there are also incompatibilities and tensions. Thus, the accountability driven goal of evaluation often sits alongside, and sometimes competes with, management and delivery logic. Evaluation is often seen by programme managers as a means of supporting improved effectiveness of implementation. In many institutional settings, funds are committed for evaluation to meet accountability purposes (Vedung, 1997), but are spent mainly for managerial, formative and developmental purposes. Nor is causal inference always absent from evaluation purposes concerned with social improvement. Nonetheless, the clusterings represented in Table 1 do summarise the main types of evaluation. These are:

(a) accountability for policy-making evaluations that rely on criteria, standards and indicators;

(b) development evaluations that adopt a change orientated approach in order to improve programmes;

(c) knowledge production evaluations that are concerned to establish causal links explanations and valid knowledge;

(d) social improvement evaluations that seek to improve the circumstances of beneficiaries by deploying change, advocacy and facilitation skills.

Already implicit in the above discussions and definitions is the dimension of time. Evaluations for the purpose of accountability tend to occur at the end of a programme cycle. Development oriented evaluations tend to occur while the programme is ongoing, and knowledge production evaluations can continue long after the initial programme cycle has ended. Wholey's concept of evaluability focuses attention on the initial programme logic while Cronbach's interest in the 'policy shaping community' carries over into the long term and the periods of transition between one programme and another. Notions of *ex-ante* evaluation (and appraisal or needs analysis), ongoing or mid-term evaluations, and *ex-post* evaluations that have been adopted as a basic framework by the European Commission and other agencies, derive from these different understandings of when evaluation activity is most relevant.

The main types of evaluation identified above can be further elaborated in terms of the kinds of questions they ask, the stakeholders that are included and the focus of their activities.

Accountability for policy-making, evaluation meets the needs of external stakeholders who require the delivery of programme or policy outputs. Management may also demand accountability but here we mean external management rather than management internal to a programme or policy area. This has become a dominant form of evaluation in public administrations, consistent with the growth of performance management philosophies more generally. Evaluations of this type tend to occur at the end of a programme or policy cycle and focus on results.

Development evaluation follows the lifecycle of an initiative with the aim of improving how it is managed and delivered. These evaluations are more likely to meet the needs of internal managers and partners rather than external stakeholders. Formative evaluations and process evaluations tend to fall into this category.

Knowledge production evaluation is mainly concerned with understanding in the longer term. These evaluations often seek to synthesise understanding coming from a number of evaluations. While both of the previous evaluation types are expected to affect current programme learning and knowledge production, this type looks to apply lessons to future programmes and policies.

Social improvement evaluation can take many forms. Many social and economic programmes depend for their success on consensus among the intended beneficiaries. Participative evaluations that seek to involve target groups contribute to the development of consensus and consent. This type of evaluation may also take on an advocacy role: promoting certain interests or groups. It is within this evaluation purpose that programme beneficiaries are most likely to be directly involved, not merely consulted.

These different evaluation types, can be further elaborated, in terms of the following questions:

(a) who are the stakeholders?

(b) what is the focus of the evaluation?

(c) what are the main approaches and methods?

(d) what are the key questions that can be asked?

Table 2 presents the main elements of the four evaluation types in relation to these questions.

However, we would not wish to suggest that these types do full justice to the diversity of evaluation models; rather they summarise the main high level differences. It is possible, for example, to see the emergence of sometimes contradictory evaluation subtypes in recent years. Two examples of these are outcome focused evaluations and participation focused evaluations.

2.5. A focus on results and outcomes

The concern that public interventions should lead to specific and measurable results is mirrored in the development of evaluation practice. In complex socioeconomic programmes in particular, the tendency is often to focus on intermediate outcomes and processes of implementation. Sometimes this is inevitable, when the final results of interventions will only be discernible in the long term. Contemporary models of public management create a demand for methods that focus on results and there has been considerable investment in such methodologies in recent years. These methodologies tend to be in three areas.

Table 2: **Evaluation types**

Purpose	**Stakeholder**	**Focus**	**Main evaluation approaches**	**Key questions**
Accountability for policy-makers	Parliaments, Ministers, funders/sponsors, Management Boards	Impacts, outcomes, achievement of targets, value for money	Indicators, performance measures, value for money studies, quantitative surveys	What have been the results? Are they intended or unintended? Are resources well-used?
Development for programme improvement	Project coordinators. Partner organisations. Programme managers	Identifying constraints. How they should be overcome? Delivery and implementation strategies.	Relating inputs to outputs, qualitative description, following processes over time	How well is the programme being managed? Can it be implemented better?
Knowledge production and explanation	Programme planners, policy-makers. Academics	Dissemination of good practice. What works? Organisational change	Experimental and quasi-experimental studies, case studies, systematic reviews and syntheses	What is being learnt? Are there lessons that can be applied elsewhere? How would we do it next time?
Social improvement and social change	Programme beneficiaries and civil society	To ensure full involvement, influence and control by citizens and affected groups.	Stakeholder involvement, participative reviews, advocacy	What is the best way to involve affected groups? How can equal opportunities and social inclusion be ensured?

The first deals with systematic reviews. Reaching policy conclusions and taking actions on the basis of the evaluation of single projects, or even programmes, has for long been criticised. The evidence-based policy movement works on the assumption that it is necessary to aggregate the results of different evaluations through systematic reviews in order to produce reliable evidence.

Next is results based management. This is now a feature of most public management systems and can be variously expressed in terms of targets, league tables, payment-by-results and outcome funding. Within the Commission there has been a move in this direction, under the label of activity based management. It is also the underlying principle of the performance reserve within the Structural Funds and relevant to current debates about impact assessment.

Finally there are macro and micro economic models. These seek to simulate the relationship between key variables and explain outputs through a mixture of available data and assumed causal relationships. Such models are especially useful where data sources are incomplete and results have to be estimated rather than precisely measured.

2.6. Participatory methods and devolved evaluations

There is a general tendency in programme and policy evaluation for multiple stakeholder and citizen involvement. These general developments have led to a spate of innovations among evaluators, who are now able to draw on an extensive repertoire of participative methods and techniques, many of them pioneered in international development contexts. They include: rapid appraisal methods, empowerment evaluation, methods for involving stakeholders, and user-focused evaluations.

Evaluation is often seen as an instrument for developing social consensus and strengthening social cohesion. The expectation is that ownership and commitment by citizens to public policy priorities will be maximised when they have also been involved in setting these priorities and evaluating the outcomes of interventions. There is a strong managerial logic within large scale decentralised programmes to use evaluation as an instrument to strengthen programme management by diffusing the culture of evaluation among all programme participants.

This also focuses attention more generally on who undertakes evaluations and where is evaluation located? Among the types outlined above, the assumption is that some outside expert occupies the evaluator role. Already within the participative subtype just referred to, the role of the evaluator is far less prominent. The role of the evaluator – as orchestrator, facilitator and enabler – is further elaborated in the discussion of constructivist evaluation in a later section of this chapter. However, even within other types of evaluation there are different possible types of operationalisations and locations of the evaluation role. One important variant is devolved evaluation.

It is becoming increasingly common for evaluation to become a devolved 'obligation' for programme beneficiaries. In the European Structural Funds, requirements for *ex-ante* and mid-term evaluations are now explicitly the responsibility of Member States and monitoring committees. The same is true of international development aid within the CEC, where project evaluation is consistently devolved to beneficiaries. Often, those who evaluate on such a 'self-help' basis are required to undertake the evaluations and must demonstrate that they incorporate and use findings. In fact, the 'devolution chain' is far more extended. In the European Structural Funds, monitoring committees will often require beneficiaries and programme managers to conduct their own evaluations. The same is true for national programmes. In the UK for example, 'local evaluation' conducted by projects within a programme are the norm. These are variously intended to focus on local concerns, inform local management and generate data that will be useful for accountability purposes.

These intentions are not conflict-free. For example, top-down demands by the EU or by central governments can easily undermine the local focus on local needs (Biott and Cook, 2000). Nonetheless, the role of devolved evaluation in the management and 'steering' of programmes has been a noticeable trend over the last ten years. By requiring programme participants to clarify their priorities, collect information, interpret findings and reflect on the implications, it is assumed that programme management at a systemic level will be improved.

2.7. Theory and practice

It has not been the intention to focus on evaluation practice in this chapter. However, it is worth reflecting briefly on how evaluation practice relates to some of the main debates outlined above:

(a) evaluations in the public sector are firmly within 'accountability' and 'programme management' purposes (e.g. Nagarajan and Vanheukelen, 1997);

(b) the notion of 'goal-free' evaluation that does not start from the objectives of programmes has never been favoured by public administrations in Europe or elsewhere. Although there is often scope to examine overall impacts and to consider 'unintentional consequences' the design of most evaluations is firmly anchored around goals and objectives;

(c) there is a trend to take on board evaluation criteria such as relevance, efficiency, effectiveness, impact and sustainability (the now standard World Bank and OECD criteria). In the EU guide referred to above, these are applied as evaluative judgements, in relation to programme objectives and in relation to socioeconomic problems as they affect target populations;

(d) there is sometimes confusion between economic appraisal and evaluation. In most public administrations judgements have to be made, before new policy initiatives are launched, on whether to proceed or not (e.g. the UK Treasury's Green Book and recent EU guidance on impact assessment). For many economists, this pre-launch appraisal is seen as the same as evaluation. In general it is sensible to confine the term evaluation to what happens once a programme or policy has been decided on;

(e) macro-economic methods in particular are more difficult to apply, when the resource input is relatively small. Where policy inputs are large scale and can be isolated from other inputs (e.g. in Objective 1, but not Objective 2 or 3 within EU Structural Funds), they can be more easily applied;

(f) the status of stakeholders has undoubtedly been enhanced in most evaluations in recent years. However, the role of stakeholders is generally as informants rather than sources of evaluative criteria, let alone as judges of merit and worth. There is considerable scope within decentralised and devolved evaluation systems for participative and constructivist approaches. What is more common is close working with stakeholder to define criteria for evaluation and contribute to consensus process;

(g) the boundaries between research and evaluation remain clouded. Many studies commissioned as evaluation contribute to knowledge production and are indistinguishable from research. Various distinctions have been proposed, including the short- rather than long-term nature of evaluation and its mainly instrumental intent. However, few of these distinctions are watertight. For example, many of the elements within this overall study could be defined as research and, arguably, once a research-generated study is deployed for evaluative purposes, its character changes.

3. The object of VET evaluation

What is evaluated is a factor in how evaluation is practised. The object of evaluation is different in different domains (health, transport, education, vocational training, etc.) and this partly shapes what we call evaluation in these various domains. At the same time, there are overarching characteristics of evaluation objects that are similar across domains. In this section, we consider both approaches to the object of evaluation; those that follow from the nature of the domain and those that follow from the characteristics of what is being evaluated.

3.1. The domain of VET

VET is a broad field that, at a minimum, includes initial vocational training, continuing vocational training, work-based learning and VET systems.

However, this selection understates the scope of VET as an object of evaluation. VET has become more complex and multi-faceted over the years as can be seen in previous Cedefop reports on vocational training research in Europe. This is mainly because there has been a shift from decontextualised studies of impact to studies that increasingly incorporate context. So VET, even at the level of the firm, is seen as being embedded in other corporate policies and procedures such as marketing, the organisation of production, supervisory and managerial practice and human resource management. In order to describe, let alone explain, the impact of VET, this broader set of factors needs to be considered. The same is true for policy level interventions. For example, what is called 'active labour-market policies', especially for those who are marginalised in the labour market, usually includes VET, but this is embedded in a raft of other policies including subsidies to employers, restructuring of benefits and new screening and matching processes.

This recontextualisation of the objects of evaluation is happening across many fields of evaluative enquiry. Evaluations of health are no longer confined to studies of illness. The 'new public health' incorporates environmental, lifestyle and policy elements alongside data on illness topics such as morbidity and mortality. Similarly, evaluation in education is now more likely to include learning processes, socioeconomic and cultural factors and broader pedagogic understandings alongside studies of classroom behaviour. It is probably more useful to think of evaluation configurations as composites of contingent evaluation objects rather than a single evaluation object.

As we shall see below, methodological developments within evaluation mirror this contextualisation. There is a move away from *ceteris paribus* assumptions, to focus increasingly on impacts in context. It is likely that the broadening conception of evaluation configurations such as VET is the result of new methods and theories helping redefine the core concept. As is often the case, methods and methodologies interact with core content, which they also help to shape.

Overall, most classes of evaluation object can be found under the umbrella of VET and it is not possible to associate the evaluation and impact of VET with a particular type of evaluation object or configuration. What is clear is that VET, as an object of evaluation, calls on a vast range of disciplinary understandings, levels of analysis and potential areas of impact. The importance of interdisciplinary evaluation efforts is highlighted by this discussion.

The scope of VET itself is further complicated by the different understandings of impact that characterise the field. We have particular studies of the impact of continuing vocational training (CVT): on company performance; on active labour markets; on individual employment and pay prospects; pedagogic methods as they influence learning-outcomes and competences; VET system reform affecting training outcomes; and knowledge and qualifications as an influence on national economic performance.

It is these clusters of interest – the preoccupations of a domain at any given time – that circumscribe the object of evaluation. It is the sets of objects and understandings around what has been called configurations that best describes what distinguishes evaluation of VET from other domains. The impacts of CVT on company performance, the way in which it is possible to improve initial vocational training through changing qualification systems, and how VET affects economic performance and social integration are all examples of what defines the object of evaluation within VET. Such preoccupations also change over time. It is worth adding that such preoccupations are also encapsulated in theoretical form. Topical theories – such as social exclusion, human capital, cultural capital, corporate innovation – will be widely accepted in the VET domain as in others. Today's theories also help define the evaluation object (see below for more general discussion of theory in evaluation).

3.2. Overarching characteristics of evaluation objects

Although it is not possible within the scope of this chapter to offer a full typology of evaluation objects, it is worth highlighting the kinds of differences that occur not only in the evaluation of VET but also in many other evaluation domains. There are many ways in which evaluation configurations can be differentiated; for simplicity's sake the following examples concentrate on common dimensions such as similarity or difference, more or less, etc. Of course, there are also much more complex descriptions of evaluation configurations.

There are a number of important dimensions of evaluation configurations, including input characteristics. Most programmes are operationalised through inputs or policy instruments; in VET these include new curricula, new forms of funding for enterprise-based training or new training courses. Such inputs may be standardised across a programme or may be more or less diverse. It is, for example, common for inputs to be carefully tailored to individual, local or enterprise needs. This will have consequences for sampling and scale of an evaluation. More seriously it will have implications for the possibilities of generalisations that can be made on the basis of evaluation findings.

Another dimension is the immediate context. The context or setting within which an input is located can also be relatively standardised or relatively diverse. This statement might apply at a spatial level (characteristics of the area) or in terms of the context of delivery or the institutional setting within which programmes are located or policies are expected to have an impact. In VET, the relevant context may be a labour market, a training provider or an enterprise. A highly diversified initiative may be located across different kinds of contexts and, even within a single context, there may be considerable variety. The diversity or standardisation of the immediate context will have many implications, in particular for how policies and programmes are implemented and how much effort needs to be devoted to the evaluation of implementation.

Modes of delivery are also important since the same input or instrument can be delivered in very different ways. For example, a needs analysis may be undertaken through a local survey as part of the recruitment process of potential trainees or by a company reanalysing its personnel data. Nowadays it would be more common for programmes to be delivered through partnership arrangements rather than through a single administrative chain. This will often be the case, for example, in VET measures delivered through EU Structural Funds.

Settings need considerations as well given the embedded and contextualised nature of many evaluation objects and that isolated evaluation objects are increasingly rare. With conceptualisations that incorporate context, evaluation objects have a tendency to become configurations. A classic example of an evaluation object that is presumed to be isolated is classroom-based studies that ignore the overall school context or the socioeconomic characteristics of a catchment area. By contrast, a VET measure that is bundled together with a package of incentives, vocational guidance measures and qualifications will need to be evaluated in this wider context.

A further dimension is the number of stakeholders. In any evaluation, there will be those who have an interest in the evaluation and what is being evaluated. Within decentralised, multi-agency programmes there are often many stakeholders, each with their own evaluation questions and judgement criteria. These might, for example, include regional authorities, training providers, sectoral representatives, social partners and European institutions.

Finally, there is the degree of consensus. Policies and programmes may be contentious and will be supported by a greater or lesser degree of consensus among stakeholders. Numerous stakeholders are often associated with lower levels of consensus. Evaluations which draw a high level of consensus will be able easily to apply agreed criteria. Where there is lower consensus, quite different criteria may need to be applied to evaluation data and more work may need to be done to bring together different interests and perspectives. This not only shapes methodology but also the work required of evaluators.

While each of these characteristics or dimensions has consequences for the design of an evaluation and how it is organised, they also interact. For example, we can envisage two different scenarios. In the first, a single subsidy is available to employers within firms in the retail sector to provide additional training to young apprentices following a recognised national qualification. In the second, a package of measures locally determined by partnerships of companies, training providers and regional authorities is available to firms, colleges and private training providers, to improve the vocational skills and work preparedness of the young unemployed.

Within the first scenario it would be possible and appropriate to assess success in terms of a limited range of output and outcome measures and possibly to apply experimental and random assignment techniques as part of the evaluation procedure. Within this scenario there would be limited resources devoted to the evaluation of the processes of implementation. There is also likely to be a limited number of stakeholders involved in the evaluation.

Within the second scenario there will be a need for several different measures of output and outcomes. Comparisons across the programme will be difficult to standardise given the diversity of modes of delivery and types of input or policy instrument. There is also likely to be limited consensus among the many different stakeholders involved in the programme and its implementation. The use of experimental methods (e.g. control groups and before and after measures) may be possible in such a configuration. It is also likely that case studies that illustrate the way all the various dimensions come together will be appropriate.

4. Philosophical foundations

4.1. Positivism, observation and theory

Before addressing particular aspects of evaluation theory it is important to locate the role of theory in evaluation within the broader set of debates within the philosophy of science. The dominant school, much criticised but of continuing influence in the way we understand the world, is logical positivism. Despite being largely discredited in academic circles for some 50 years, this school still holds sway in policy debates. It constitutes the base model around which variants are positioned. With a history that stretches back to Compte, Hume, Locke, Hobbes and Mill, positivism emerged partly as a reaction to metaphysical explanations: that there was an 'essence' of a phenomenon that could be distinguished from its appearance. At the heart of positivism therefore, is a belief that it is possible to obtain objective knowledge through observation and that such knowledge is verified by statements about the circumstances in which such knowledge is true.

In the field of evaluation, House (1983) has discussed this tradition under the label of objectivism: 'Evaluation information is considered to be "scientifically objective." This objectivity is achieved by using "objective" instruments like tests or questionnaires. Presumably, results produced with these instruments are reproducible. The data are analysed by quantitative techniques which are also "objective" in the sense that they can be verified by logical inspection regardless of who uses the techniques.' (House, 1983; p. 51).

House goes on to emphasise that part of objectivist tradition that he calls 'methodological individualism' in Mill's work in particular. Thus, repeated observation of individual phenomena is the way to identify uniformity within a category of phenomena. This is one important strand in the mainstream of explanations within the social and economic sciences. It is the basis for reductionism: the belief that it is possible to understand the whole by investigating its constituent parts.

'By methodological individualism, I mean whatever methodologically useful doctrine is asserted in the vague claim that social explanations should be ultimately reducible to explanations in terms of people's beliefs, dispositions, and situations. […] It is a working doctrine of most economists, political scientists, and political historians in North America and Britain.' (Miller, 1991; p. 749).

In this world-view, explanations rest on the aggregation of individual elements and their behaviours and interactions. It is worth noting that this has been described as a 'doctrine' as well as a methodological statement. It underpins many of the survey based and economic models that are used in evaluation.

There is now widespread agreement that empirical work cannot rely only on observations. There are difficulties empirically observing the entirety of any phenomena; all description is partial and incomplete, with important unobservable elements. 'Scientists must be understood as engaged in a metaphysical project whose very rules are irretrievably determined by theoretical conceptions regarding largely unobservable phenomena.' (Boyd, 1991; p. 12). This is even more true for mechanisms which it is generally recognised can be imputed but not observed. As Boyd goes on to say, 'it is an important fact, now universally accepted, that many or all of the central methods of science are theory dependent'.

This recognition of the theory dependence of all scientific inquiry underpins the now familiar critiques of logical positivism, even though there is considerable difference between the alternatives that the critics of positivism advocate.

The two most familiar critiques of positivism are scientific realism and constructivism.

4.2. Scientific realism

Scientific realism, while acknowledging the limits of what we can know about phenomena, asserts that theory describes real features of a not fully observable world. Not all realists are the same and the European tradition currently being inspired mainly by the work of Pawson (Pawson and Tilley, 1997; Pawson, 2002a and b) can be distinguished in various ways from US realist thinking. For example, some prominent North American realists commenting on Pawson and Tilley's work have questioned the extent to which realists need completely to reject experimental and quasi-experimental designs, and suggest that more attention should be paid in the realist project to values. This is especially important if, in addition to explanation, realists are to influence decisions (Julnes et al., 1998). Nonetheless, this chapter draws mainly on the work of Pawson and Tilley to describe the realist position in evaluation.

In some ways realism continues the positivist project: it too seeks explanation and believes in the possibility of accumulating reliable knowledge about the real world, albeit through different methodological spectacles. According to Pawson and Tilley, it seeks to open the 'black-box' within programmes or policies to uncover the mechanisms that account for what brings about change. It does so by situating such mechanisms in contexts and attributing to contexts the key to what makes mechanisms work or not work. This is especially important in domains such as VET where the evaluation objects are varied and drawn from different elements into different configurations in differentiated contexts.

'What we want to resist here is the notion that programs are targeted at subjects and that as a consequence program efficacy is simply a matter of changing the individual subject.' (Pawson and Tilley, 1997; p. 64).

Rather than accept a logic that sees programmes and policies as simple chains of cause and effect, they are better seen as embedded in multilayered (or stratified) social and organisational processes. Evaluators need to focus on 'underlying mechanisms': those decisions or actions that lead to change, which is embedded in a broader social reality. However these mechanisms are not uniform or consistent even within a single programme. Different mechanisms come into play in different contexts, which is why some programmes or policy instruments work in some, but not all, situations.

Like all those interested in causal inference, realists are also interested in making sense of patterns or regularities. These are not seen at the level of some programme level aggregation but rather at the underlying level where mechanisms operate. As Pawson and Tilley (1997; p. 71) note: 'regularity = mechanism + context'. Outcomes are the results of mechanisms unleashed by particular programmes. It is the mechanisms that bring about change and any programme will probably rely on more than one mechanism, not all of which may be evident to programme architects or policy-makers.

As Pawson and Tilley summarise the logic of realist explanation: 'The basic task of social inquiry is to explain interesting, puzzling, socially significant regularities (R). Explanation takes the form of positing some underlying mechanism (M) which generates the regularity and thus consists of propositions about how the interplay between structure and agency has constituted the regularity. Within realist investigation there is also investigation of how the workings of such mechanisms are contingent and conditional, and thus only fired in particular local, historical or institutional contexts (C)' (Pawson and Tilley, 1997; p. 71).

Applying this logic to VET, we may note, for example, that subsidies to increase work-based learning and CVT in firms sometimes lead to greater uptake by the intended beneficiaries. This need not lead to the assessment of the programme as ineffective because, for example, positive outcomes can only be observed in 30 % of cases. We try rather to understand the mechanisms and contexts which lead to success. Is the context one where firms showing positive outcomes are in a particular sector or value chain or type of region? Or is it more to do with the skill composition of the firms concerned? Are the mechanisms that work in these contexts effective because a previous investment has been made in work-based learning at the firm level or is it because of the local or regional training infrastructure? Which mechanisms are at play and in what context:

(a) the competitive instincts of managers (mechanism), who fear that their competitors will benefit (context) unless they also increase their CVT efforts?

(b) the demands of trade unions concerned about the professionalisation and labour-market strength of their members (mechanism), sparked off by their awareness of the availability of subsidies (context)?

(c) the increased effectiveness of the marketing efforts of training providers (mechanism) made possible by the subsidies they have received (context)?

According to the realists, it is by examining and comparing the mechanisms and contexts in which they operate in relation to observed outcomes, that it becomes possible to understand success and describe it. For Pawson and Tilley, all revolves around these CMO (context, mechanism, outcome) configurations.

Policy-makers are then in a position to consider options such as:

(a) focusing the programme more narrowly at beneficiaries that are likely to change because of the mechanisms that work in the contexts they inhabit;

(b) differentiating a programme and its instruments more clearly to ensure that different mechanisms that work in different contexts are adequately covered;

(c) seeking to influence the contexts within which the programme aims to be effective.

The table below, taken from the concluding chapter of Pawson and Tilley's book, provides a brief summary of the realist position, in terms of eight 'rules' that are seen as encapsulating the key ideas of realistic enquiry and method.

Table 3: **Rules guiding realistic enquiry and method**

Rule 1:	Generative causation Evaluators need to attend to how and why social programmes have the potential to cause change
Rule 2:	Ontological depth Evaluators need to penetrate beneath the surface of observable inputs and outputs of a programme
Rule 3:	Mechanisms Evaluators need to focus on how the causal mechanisms which generate social and behavioural problems are removed or countered through the alternative causal mechanisms introduced in a social programme
Rule 4:	Contexts Evaluators need to understand the contexts within which problem mechanisms are activated and in which programme mechanisms can be successfully fired
Rule 5:	Outcomes Evaluators need to understand what are the outcomes of an initiative and how they are produced
Rule 6:	CMO configurations In order to develop transferable and cumulative lessons from research, evaluators need to orient their thinking to context-mechanism-outcome pattern configurations (CMO configurations)
Rule 7:	Teacher-learner processes In order to construct and test context-mechanism-outcome pattern explanations, evaluators need to engage in a teacher-learner relationship with program policy-makers, practitioners and participants
Rule 8:	Open systems Evaluators need to acknowledge that programmes are implemented in a changing and permeable social world, and that programme effectiveness may thus be subverted or enhanced through the unanticipated intrusion of new contexts and new causal powersAdapted from Pawson and Tilley (1997)

Source: Adapted from Pawson and Tilley (1997)

4.3. Constructivists

Constructivists deny the possibility of objective knowledge about the world. They follow more in the tradition of Kant and other continental European philosophers than the mainly Anglo Saxon school that underpins positivism and realism. It is only through the theorisations of the observer that the world can be understood.

'Socially constructed causal and metaphysical phenomena are, according to the constructivist, real. They are as real as anything scientists can study ever gets. The impression that there is some sort of socially unconstructed reality that is somehow deeper than the socially constructed variety rests, the constructivist maintains, on a failure to appreciate the theory-dependence of all our methods. The only sort of reality any of our methods are good for studying is a theory-dependent reality.' (Boyd, 1991; p. 13).

The way we know, whatever the instruments and methods we use, is constructed by human actors or stakeholders. According to Stufflebeam in his review of *Foundation models for 21st century program evaluation*: 'Constructivism rejects the existence of any ultimate reality and employs a subjectivist epistemology. It sees knowledge gained as one or more human constructions, uncertifiable, and constantly problematic and changing. It places the evaluators and program stakeholder at the centre of the inquiry process, employing all of them as the evaluation's "human instruments". The approach insists that evaluators be totally ethical in respecting and advocating for all the participants, especially the disenfranchised.' (Stufflebeam, 2000a; pp. 71-72).

The most articulate advocates of constructivism in evaluation are Guba and Lincoln. They have mapped out the main differences between constructivists and the 'conventional' position (as they label positivists) in their well-known text *Fourth generation evaluation* (Guba and Lincoln, 1989). The highlights of this comparison is summarised in the table below:

Table 4: **Comparing constructivist and 'conventional' evaluation**

	Conventional	Constructivist
Nature of truth	The truth of any proposition (its factual quality) can be determined by testing it empirically in the natural world. Any proposition that has withstood such a test is true; such truth is absolute	The truth of any proposition (its credibility) can be determined by submitting it semiotically to the judgement of a group of informed and sophisticated holders of what may be different constructions. Any proposition that has achieved consensus through such a test is regarded as true until reconstructed in the light of more information or increased sophistication; any truth is relative.
Limits of truth	A proposition that has not been tested empirically cannot be known to be true. Likewise, a proposition incapable of empirical test can never be confirmed to be true.	A proposition is neither tested nor untested. It can only be known to be true (credible) in relation to and in terms of informed and sophisticated constructions.
Measurability	Whatever exists in some measurable amount. If it cannot be measured it does not exist.	Constructions exist only in the minds of constructors and typically cannot be divided into measurable entities. If something can be measured, the measurement may fit into some constructions but it is likely, at best, to play a supportive role.
Independence of facts and theories	Facts are aspects of the natural world that do not depend on theories that happen to guide any given inquiry. Observational and theoretical languages are independent.	Facts are always theory-laden, that is, they have no independent meaning except within some theoretical framework. There can be no separate observational and theoretical languages.
Independence of facts and values	Facts and values are independent. Facts can be uncovered and arrayed independently of the values that may later be brought to bear to interpret or give meaning to them. There are separate factual and valuational languages, the former describing 'isness' and the latter 'oughtness'.	Facts and values are interdependent. Facts have no meaning except within some value framework; they are value-laden. There can be no separate observational and valuational languages.

Source: adapted from Guba and Lincoln, 1989

According to Guba and Lincoln, when considering the purpose of evaluations, one needs to distinguish both between merit and worth and between summative and formative intent:

(a) a formative merit evaluation is one concerned with assessing the intrinsic value of some evaluand with the intent of improving it; so, for example, a proposed new curriculum could be assessed for modernity, integrity, continuity, sequence, and so on, for the sake of discovering ways in which those characteristics might be improved;

(b) a formative worth evaluation is one concerned with assessing the extrinsic value of some evaluand with the intent of improving it; so, for example, a proposed new curriculum could be assessed for the extent to which desired outcomes are produced in some actual context of application, for the sake of discovering ways in which its performance might be improved;

(c) a summative merit evaluation is one concerned with assessing the intrinsic value of some evaluand with the intent of determining whether it meets some minimal (or normative or optimal) standard for modernity, integrity, and so on. A positive evaluation results in the evaluand being warranted as meeting its internal design specifications;

(d) a summative worth evaluation is one concerned with assessing the extrinsic value of some evaluand for use in some actual context of application. A positive evaluation results in the evaluand being warranted for use in that context. (Guba and Lincoln, 1989; pp. 189-190).

In practical terms, what it is that the evaluator should do, Guba and Lincoln start from the 'claims, concerns and issues' that are identified by stakeholders, people 'who are put at some risk by the evaluation'. It is therefore necessary for evaluators to be 'responsive'. 'One of the major tasks for the evaluator is to conduct the evaluation in such a way that each group must confront and deal with the constructions of all others, a process we shall refer to as hermeneutic dialectic. […] Ideally responsive evaluation seeks to reach consensus on all claims, concerns and issues […]' (Guba and Lincoln, 1989; p. 41).

A distinctive role of the evaluator, therefore, is to help put together 'hermeneutic' circles. This is defined by Guba and Lincoln as a process that brings together divergent views and seeks to interpret and synthesise them mainly to 'allow their mutual exploration by all parties' (Guba and Lincoln, 1989; p. 149). As Schwandt has argued from a postmodernist standpoint, 'only through situated use in discursive practices or language games do human actions acquire meaning' (Schwandt, 1997; p. 69). Applied to evaluation, this position argues for the importance of the 'dialogic encounter' in which evaluators are 'becoming partners in an ethically informed, reasoned conversation about essentially contested concepts […]' (Schwandt, 1997; p. 79).

In more down to earth terms, Guba and Lincoln emphasise the role of the evaluator to:

(a) prioritise those unresolved claims, concerns and issues of stakeholders that have survived earlier rounds of dialogue orchestrated by the evaluator;

(b) collect information through a variety of means – collating the results of other evaluations, reanalysing the information previously generated in dialogue among stakeholders, conducting further studies – that may lead to the 'reconstruction' of understandings among stakeholders;

(c) prepare and carry out negotiations that, as far as possible and within the resources available, resolve that which can be resolved and (possibly) identify new issues that the stakeholders wish to take further in another evaluation round.

So how might this be exemplified in the VET domain? It should be noted that what follows does not fully conform to Guba and Lincoln's vision of constructivist evaluation, largely because it is situated in a larger scale socioeconomic policy context than many of their own smaller scale case examples. But also it should be noted that constructivist thinking is, to some extent, relevant to many contemporary evaluation challenges and the example below is intended to illustrate such potential relevance.

So, to apply this logic to VET, constructivist thinking can be especially helpful where there is a problem area with many stakeholders and the entire system will only be able to progress if there is a broad consensus. For example, there may be a political desire to become more inclusive and involve previously marginalised groups in training opportunities. The problem is how to ensure that certain groups such as women, young people and ethnic communities are given a higher profile in VET. Here, the involvement of many stakeholders will be inevitable. Furthermore the views of these stakeholders are more than data for the evaluator: they are the determinants and shapers of possible action and change. Unless the trainers, employers, advocacy groups, funding authorities and employment services responsible for job-matching and the groups being 'targeted' cooperate, change will not occur. It is also likely that these stakeholders hold vital information and insights into the past experience of similar efforts; what went wrong and right and what could be done to bring about improvements in the future.

The evaluator might then follow much of the constructivist logic outlined above:

(a) identify the different stakeholders who potentially have a stake in these areas of concern;

(b) conduct a series of initial discussions to clarify what they know, what they want and what are their interests;

(c) feed back to all stakeholders their own and each other's interests, knowledge and concerns in a way that emphasis the similarities and differences;

(d) clarify areas of agreement and disagreement and initiate discussions among the stakeholders and their representatives to clarify areas of consensus and continuing dissent;

(e) agree what other sources of information could help move the stakeholders forward – perhaps by synthesising other available studies, perhaps by initiating new studies;

(f) reach the best possible consensus about what should be done to improve VET provision and participation for the groups concerned.

It is worth highlighting that the balance of activities within constructivist evaluation is very different from both positivist and realist variants. It emphasises the responsive, interactive, dialogic and 'orchestrating' role of the evaluator because the sources of data that are privileged are seen to reside with stakeholders, as much as with new studies and externally generated data.

5. Evaluation theory

There has been a strong bias within evaluation as it has evolved to focus on method, technique and, to a lesser extent, methodology. It is only in recent years that there has been an upsurge in interest in the role of theory in evaluation. To some extent, this reflects the wider debates from within the philosophy of science that has been sketched out above. From the early 1990s onwards there has been a re-balancing of attention towards theory. Chen's book, *Theory-driven evaluations* (1990), has become a landmark in this shift in focus towards theory. The now classic text *Foundations of evaluation* (Shadish et al., 1991) is organised around five main bodies of theory: social programming, knowledge construction, valuing, knowledge use and evaluation practice.

As it has been widely recognised, Weiss was among the first to direct our attention to the importance of theory (Weiss, 1972) and has actively carried forward this debate under the umbrella of the Aspen Institute's *New approaches to evaluating Community initiatives* (Connel et al., 1995; Fulbright-Anderson et al., 1998). While the starting point of the discussion that follows is these authors, it continues to need to be situated in the broader philosophical debates outlined earlier.

5.1. Programme theory

The dominant school of theory in evaluation is 'programme theory'. This is concerned with opening up the programme 'black-box', going beyond input/output descriptions and seeking to understand how programmes do and do not work.

Chen's conceptualisation distinguishes 'normative' and 'causative' components of programme theory which he defines as 'a specification of what must be done to achieve the desired goals, what other important impacts may also be anticipated and how these goals and impacts would be generated' (Chen, 1990; p. 43).

Chen's conceptualisation extends to what he identifies as 'six domains'.

'The following three domain theories are part of the general normative theory: 1) treatment theory specifies what the nature of the program treatment should be; 2) implementation environment theory specifies the nature of the contextual environment within which the program should be implemented; 3) outcomes theory specifies what the nature of the program outcomes should be.

The following three domain theories are related to the general causative theory: 1) impact theory specifies the causal effect between the treatment and the outcome; 2) intervening mechanism theory specifies how the underlying intervening processes operate; 3) generalization theory specifies the generalizability of evaluation results to the topics or circumstances of interest to stakeholders.' (Chen, 1990; pp. 49 and 51)

Although avowedly seeking to escape from the limitation of input/output thinking, Chen's conceptualisation is still linear. His domains follow the treatment/implementation/outcome logic and incorporate concepts such as intervening mechanisms: 'the causal processes underlying a program so that the reasons a programme does or does not work can be understood' (Chen, 1990; p. 191).

The underlying logic of causality favoured by Chen is essentially consistent with classic experimental thinking. For example, with regard to a programme that used comic books to influence adolescent smoking: 'The underlying causal mechanism of this program is the assumption that the comic book will attract adolescents' interest and attention and that they will read it closely and frequently, and thereby pick up the important anti-smoking message contained in it. The message will in turn change their attitudes, beliefs, and behaviour regarding smoking. The causal structure of this program is that the program treatment variable (exposure to the comic book) attempts to affect the intervening variable (the intensity of reading), which in turn will affect the outcome variables (attitudes, beliefs, and behaviour toward smoking)' (Chen, 1990, p. 193).

In comparison to realist evaluation approaches (see above), there continues to be an emphasis on the programme's interventions rather than on the mechanism operating within the context. For example, if we try to answer the question 'Why are some adolescents more likely to be influenced by such exposure?', answers do not fall out easily from traditional programme theory logic. It is these 'underlying mechanisms' that are not explained in this framework. Nor are the contexts within which these interventions occur explored in detail.

5.2. Theory based evaluation and theories of change

The other main strand of theory in evaluation is labelled 'theory based evaluation' and latterly 'theory of change' and is associated with the Aspen round table on comprehensive community initiatives. In Volume 1 of the Aspen collection, Weiss identifies four main rationales for theory based evaluation: '1) It concentrates evaluation attention and resources on key aspects of the program; 2) it facilitates aggregation of evaluation results into a broader base of theoretical and program knowledge; 3) it asks program practitioners to make their assumptions explicit and to reach consensus with their colleagues about what they are trying to do and why; 4) evaluations that address the theoretical assumptions embedded in programs may have more influence on both policy and popular opinion.' (Weiss, 1995; p. 69).

This is also, in essence, a programme theory approach. In a subsequent volume, Connel and Kubisch define this evaluation approach 'as a systematic and cumulative study of the links between activities, outcomes, and contexts of the initiative.' (Fulbright-Anderson et al., 1998; p. 16). They emphasise collaborative working with stakeholders to bring to the surface underlying mechanisms.

Weiss herself takes these ideas of joint working further: 'Three big advantages for pursuing a theory of change evaluation are as follows:

(a) a theory of change evaluation allows evaluators to give early word of events without having to wait until the end of the whole program sequence;

(b) the evaluators can identify which assumptions are working out and which are not. They can pinpoint where in the theory the assumptions break down. This should enable the program to take corrective action before too much time goes by;

(c) The results of a theory of change evaluation can be more readily generalized across programs. Seeing the successes and the failures between closely linked assumptions, such as between greater parental attention to children and better child behaviour, is easier than between say, parenting education programs and better child behaviour.' (Weiss, 2000).

By focusing on the assumptions of 'programme practitioners' and aspiring to encourage consensus among them, this vision of evaluation theory shares many features with the participative and even constructivist schools of evaluation. As described by some of its main proponents, the 'theory of change' evaluation approach – as it has come to be called – is also a collaborative, dialogic process. It takes the themes of practitioners, makes them coherent, explicit and testable and then seeks to measure and describe programme outcomes in these terms.

Overall, over the last seven or eight years in particular, there has been a gradual blurring of what is meant by programme theory. The term now seems to encompass several approaches that unpick the logic of programmes, make explicit their assumptions, work with stakeholders, monitor progress and explain the outcomes that are observed (Rogers, 2000 as an example of this tendency). As such, it has come to include what is now classic programme theory with theory based evaluations and realistic evaluations such as those advocated by Pawson and Tilley (Rogers, 2000; p. 219).

The programme theory approach has also been taken on board by those who advocate logic models or logical frameworks that link outcomes with programme activities and processes. Thus, a recent W.K. Kellogg Foundation guide (2000) also makes the link with 'theoretical assumptions/principles of the programme' and devotes an entire chapter to 'developing a theory of change logic model for your programme'. This is an important development given the power of logic models in the world of evaluation. These were initiated by the World Bank and taken on also by the EU. One of the main criticisms of these models is their lack of explanatory power and a-theoretical nature. Bringing theory-based approaches into the logic model framework begins to address some of these criticisms.

5.3. A wider theoretical frame

However, we would not wish to confine descriptions of theory solely to programme theory and associated elaborations. The Shadish et al. (1991) framework includes a theory of social programming. However, their focus is more on a theory of what social programmes do and how effective they are. This is consistent with their overall approach which is to elaborate the theoretical basis for evaluation practice. Other theoretical focuses for Shadish et al. concern:

(a) the theory of use, i.e. what is known about how to encourage use;

(b) the theory of valuing, i.e. about judging outcomes and the role of values and stakeholder interests in such judgements;

(c) the theory of knowledge in evaluation, i.e. the familiar questions of what constitutes knowledge, explanation and valid method;

(d) the theory of practice in evaluation, i.e. the main decisions about resource allocation, the choice of methods, questions to ask and evaluation purposes.

Beyond the various focuses identified – that is programme theory and its various elaborations – and theories of evaluation itself, as articulated by Shadish et al., there are a number of other bodies of theory that are undoubtedly relevant and come into the discourse of evaluators. In particular, there are domain theories and implementation theories and change.

In every policy domain or field where evaluation occurs, there are bodies of theory unrelated to the practice of evaluation and to the logic of programmes. Thus, in social welfare, there are theories related to the welfare state, the nature of social solidarity, the behavioural consequences of different benefit regimes and the interactions between social welfare and labour-market performance. Similar bodies of theory exist in relation to VET as they do in other domains such as research and development, regional planning, education and criminal justice. In the European context, at least, there seems to be an expectation that evaluators will have some knowledge of the domain contexts within which they work, including relevant domain theories.

There are also substantial bodies of relevant theory about policy change and implementation deriving mainly from political science and policy studies. For example this extends beyond Chen's description of implementation environment evaluation (Chen, 1990). It is well exemplified by the work of people such as Sabatier (1988) and Sabatier and Jenkins-Smith (1993). There is also more generic literature on implementation and change, often encapsulated under the heading 'the diffusion of innovation' and following on from the work of Rogers (1995). This is particularly relevant to the issue of generalisability of innovations that are first broached on pilot basis.

In summary, five bodies of theory appear to be relevant to evaluators:

(a) theories of evaluation. These would include programme theory; theories of change approaches and realist approaches which emphasise the identification of mechanisms underlying successful change which have to be understood in specific contexts and settings;

(b) theories about evaluation. Thus there is a growing literature on evaluation practice, use, design and capacity. Included in this category would be particular aspects of practice identified by Shadish et al. such as theories of valuing;

(c) theories of knowledge, including the main debates about the nature of knowledge, epistemology, methodology, etc., and about the nature of causal inference;

(d) domain and thematic theories, which could be described as theory of the evaluation object. This would include bodies of theory about domains such as human resource development, skill acquisition, the development of human capital and equal opportunities that could inform evaluation design, programme/policy implementation and outcomes;

(e) theories of implementation and change often seen as relevant by evaluators. We would include here understandings of policy change, the diffusion of innovation and administrative behaviour. Such bodies of theory are likely to condition the success of programme interventions and can be quite separate from the kind of programme theories referred to above.

Finally, it is worth restating the main reasons that theory is seen as important in evaluation:

(a) theory can help support interpretations. This follows from the widespread recognition that all evaluations are based on data that can be interpreted in different ways. Theory provides an explicit framework for such an interpretation;

(b) theory can help fill in the gaps in incomplete data. This follows from the recognition that however thorough evaluators may be, they will never have the complete picture. Theory can provide a plausible way of filling gaps in available evaluation data;

(c) theory can provide a framework to work with stakeholders. This follows from the increasingly common practice of dialogue and collaboration between evaluators, presumed beneficiaries and others who are affected by programmes and policy instruments;

(d) theory can help prediction and explanation. This is the classic scientific role of theory: to suggest and explain causal links and likely outcomes;

(e) theory can make explicit the constructed objects of evaluation. Many contemporary objects of evaluation are constructed. They are abstracted ideas which do not have a direct empirical referent. (e.g. efficiency, a learning organisation, an enterprise culture would all be examples of constructed objects.) In order to describe and measure these objects, theory is needed.

At a different level, we are also beginning to see theoretical development around the issue of complexity in socioeconomic programmes (Sanderson, 2000). Many interventions are not self-contained, they interact with other programmes and with other social and organisational processes. Thus in VET, a new training system is embedded in an institutional and educational context which supports and constrains this system. Similarly, a training initiative at the level of a particular work group is mediated by the way work is organised, different management styles and labour-market behaviour of employers and workers. Often there are multiple programmes and interventions operating simultaneously on a particular target group or area. The interaction of these various processes and programmes is one of the greatest challenges that evaluators face. Complexity theory is a very new entrant into evaluation thinking. Questions are raised more often than answers are provided.

6. Determinants of evaluation use

Evaluation types and philosophies are shaped by contexts of use. This term encompasses both factors that encourage use and what is sometimes called 'evaluation capacity development'. These contexts and capacities are not only conceived of differently by different scholars but also change as context changes. It is these kinds of use and capacity considerations that ultimately decide how evaluation contributes to policy-making and the delivery of public policies.

6.1. Instrumental versus cumulative use

There has been debate about the use of evaluation for over 25 years. This is primarily associated with two main scholars: Weiss (1976) and Patton (1997, 2002 and earlier editions). In the mid-1970s Weiss began to focus on evaluation use, criticising simplistic notions of instrumental use. Based on empirical studies of how policy-makers use evaluation, Weiss has been associated with a complex understanding of decision-making and policy-making in which evaluation findings are internalised, selectively used and rarely lead directly to specific decisions or changes in policy. This theory, which is sometimes labelled 'an enlightenment view of evaluation', takes a long-term incremental view about the way evaluation findings feed through to policy-making. In this regard, she is close to Rossi's understanding of the conceptual use of evaluation i.e. 'the use of evaluations to influence thinking about issues in a general way' (Rossi et al., 1999). She challenges the rational model of decision-making and bases her conclusions on studies of how organisations actually work. She effectively argues for cumulative learning across many evaluations rather than direct use from particular evaluations. In some of her more recent work (Weiss, 1999) she associates herself with arguments of Sabatier and Jenkins-Smith (1993) about the importance of policy forums that bring together academics and policy-makers and can act as a vehicle for the absorption of evaluation findings. She nonetheless remains true to her earlier arguments that evaluators should never expect their inputs to override political agendas and administrative necessities which may push decisions in quite different directions from their recommendations.

Patton is more committed to the instrumental purposes of evaluation. 'By utilization I mean intended use by intended users' (Patton 1997, 2002 and earlier editions). He places considerable emphasis on understanding the priorities of decision-makers, engaging with them, encouraging them to own evaluations and their results in order to enhance use. Furthermore, Patton tends to emphasise the example of individual decision-makers, particular people and not the decision itself. Indeed he is interested in individual 'decision-makers' cognitive style and logic'. Unlike Weiss, he is less interested in organisational and administrative processes and more with the act of decision and the particular decision-maker.

It has been commonly observed (e.g. Alkin, 1990) that an important determinant of the difference in perspective between Weiss and Patton is their respective fields of practice and study. Patton has worked mainly in local community and voluntary organisations; where he has worked at public administrations they have also have been at a local level. Weiss has been more occupied with large-scale national programmes (beginning for example with the 1970s' war on poverty in the US). This probably goes a long way to explaining the one's preoccupation with complex organisational processes and the other's preoccupation with individual decision-makers and their decisions.

6.2. Process use of evaluation

Although the above summarises, in a simplified way, the main debate between Weiss and Patton, both have contributed far more to our understanding of evaluation use. In particular Patton is probably the originator of the term 'process use', i.e. 'changes […] that occur among those involved in evaluation as a result of learning that occurs during the evaluation process' (Patton, 1997). Considerations of the process aspects of evaluation open up discussions also of different, more action orientated, views of evaluation purpose. For example, Patton also discusses intervention orientated evaluation, participatory evaluation and empowerment evaluation. These kinds of evaluation approaches are less relevant to the immediate concerns of this particular study, but do raise interesting methodological questions about how, in the course of evaluations, one can increase commitment and ownership by attending to the details of the evaluation process.

6.3. The importance of methodology

Weiss, especially in her early work, opens up another aspect of evaluation use, the importance of sound methodology and reliable data (Weiss and Bucuvalas, 1980). Indeed in some of this early work Weiss anticipates some very contemporary discussions about evidence-based policy making and the use of experimental methods. It should be acknowledged, however, that in later years as she investigated more carefully the actual uses made of evaluation, she shifted towards the position outlined above, that we had to understand more about use in an organisational context. In her concern for sound methods and valid research, Weiss is more 'hands on' than many of her contemporaries: Campbell, for example, sees use as being essentially part of politics and of little concern for evaluators who should be driven by 'truth'.

In many ways we can perceive a pendulum swing, between organisational and implementation concerns on the one hand and methodological and quality assurance concerns on the other, back and forth over the last 25 years. The current interest in evidence-based policy, the rediscovery of 'the gold standard' of experimental methods and randomised control trials (RCTs) and the 'realistic critique' of such trials and experiments, represents a renewed movement toward the methodological end of the pendulum. Thus the 'What works?' school (Davies et al., 2000) is pre-eminently concerned with the nature of research evidence and the need for rigorous methodology. The adherents of this school often regard RCTs as the most reliable form of primary evidence and systematic reviews and meta-analysis – which make quantitative estimates of the overall effects of interventions by aggregating primary studies – as the most reliable approach for secondary evidence. However, already in Davies et al. (2000) the editors of this collection themselves begin to address questions of better understanding the policy process. This is taken further by Nutley et al. (2003) when they move beyond data and findings to consider how evidence is disseminated and incorporated into practice. Thus these authors also follow the pendulum between methodological and implementation concerns.

However, the central components of the evidence-based policy debate remain methodological, as in the debate around meta-analysis and systematic reviews (Gallo, 1978; Pawson, 2002a). The core of this debate is the importance of drawing together evidence from previous related initiatives before embarking on new initiatives. For some, narrative reviews that are essentially structured qualitative literature reviews drawing together lessons, is their preferred approach. Others seek to quantify precise effects by looking to quantitative and experimental studies along the lines of RCTs in medical drug trials. Of course, the possibility of undertaking such quantitative studies is limited to a subset of policy areas usually in the social welfare, social benefit and welfare to work areas. Despite the methodological disagreements among the proponents of different approaches to such secondary analysis, there is a widespread consensus that bringing together evidence from across many evaluations is an important guarantor of the reliability of findings.

6.4. Evaluation and learning

The most recent attempt to revisit the issue of evaluation use in a comprehensive fashion, is a publication of the American Evaluation Association (AEA), *The expanding scope of evaluation use* (Caracelli and Preskill, 2000).

In Chapter 2, Preskill and Torres consider 'the learning dimension of evaluation use'. These authors highlight, in particular, learning at an organisational level that occurs as part of the evaluation process. They draw on constructivist theories of learning, in which learners themselves engage in an active process of interpretation and dialogue in order to construct their own meanings. They link this perspective with Lave and Wenger's (1991) notions of 'communities of practice', i.e. groups of individuals working together, who are interdependent and whose tacit knowledge and problem-solving capacities are integrated into their social and professional life. The authors suggest that learning from evaluations 'will most likely occur when evaluation is collaborative, is grounded in constructivist and transformational learning theories, and builds communities of evaluation practice.' The implication of this argument for evaluation use is to reinforce the importance of developing communities of both evaluators and users within organisations, if evaluation is to become part of an active learning process. It also steers us away from the narrow view of intended use that certainly informed the earlier work of Patton. Given that communities of practice will interpret and construct their own meanings using data and findings that they bring into their own context, the use of evaluation may not be as evaluators or the commissioners of evaluation originally intended. However, transforming such evaluative outputs and processes into their own organisational context still constitutes evaluation use.

6.5. The institutionalisation of evaluation

There is extensive literature around the role of evaluation in policy systems. This literature is largely based in policy analysis and public administration rather than in evaluation *per se*. The preoccupations of this body of literature can be seen as mainly about evaluation capacity, including: organisational learning in the public sector; how to build evaluation into public administrations; the nature of effective evaluation capacity; institutionalising and making evaluation more professional; and analyses of the policy process itself. This latter set of preoccupations is the closest 'cousin' to the earlier discussion of evaluation use. However, within this literature the starting point is the nature of the policy process which then moves on to considerations of evaluation practice. Within the earlier cited literature, the starting point is usually evaluation practice, which then moves on to considerations of policy-making.

Organisational learning has become a common metaphor within studies of various kinds of organisations, although most of this work has centred on the private sector. Leeuw et al. (1994) bring together a range of experiences from the public sector across Canada, the Netherlands, Norway, Sweden and the US that seeks to apply the concept of organisational learning through evaluation to public sector institutions. Among the themes highlighted in this study is the role of internal evaluations within public bodies (see especially chapters by Sonnichsen and by Mayne). The authors highlight the important role that can be performed by internal evaluation offices and how these can influence the overall organisations' evaluation practice and learning. They suggest that developing a 'double-loop learning' process – reporting not only to programme managers but also to top management within government departments – as an important part of the contribution that evaluation units can make. However, these studies also suggest that, while internal units can make an important contribution to organisational learning, they depend ultimately on what Sonnichsen calls 'a disposition towards critical self-examination'. His notion that 'self-reflection is crucial before organisations begin to learn' highlights the importance of creating a general evaluation culture within an organisation.

In the concluding chapter of this collection, Rist identifies two sets of preconditions for learning from evaluation. The first set arises from the importance of fitting in with the policy cycle. This recognises that organisations need 'information at different phases of the policy cycle'. Synchronising evaluation outputs with different policy needs over time is seen as an important means of encouraging learning. The second set of preconditions for learning emphasises how information is transmitted and filtered within public organisations. Thus studies within this framework suggest that 'governmental organisations appear more receptive to information produced internally than that which comes from external sources'. This is especially so when the news is bad! Bad news is easier to receive from internal rather than external evaluators. Another precondition appears to be the credibility of the sources of information. This may, on occasion, favour internally generated evaluation findings, but may also depend on who are the sponsors or gatekeepers who bring information in to a public body. This can be seen as related to wider issues of relationship-building and trust. The existence of such organisational attributes seems an important precondition for receptivity to evaluation findings and to organisational learning within the public sector.

Another collection we have considered, focuses directly on evaluation capacity (Boyle and Lemaire, 1999). Although some of the authors in this volume overlap with the Leeuw et al. collection referred to above, the concerns here are broader. Building evaluation capacity is seen in terms of 'national evaluation systems'. Thus, evaluation capacity goes beyond the internal organisation of public bodies to include the location of evaluation in the executive or the legislature (e.g. Parliament) and broader issues of governance and institutional arrangements. For our purposes a number of themes explored in this collection are particularly relevant.

One such theme is the design of evaluation functions and offices. Sonnichsen considers the advantages and disadvantages of a centralised and a decentralised model of evaluation functions. He notes the potential for greater independence and credibility of centralised functions and their ability to develop strategic evaluation plans. This model has been favoured in Canada (for example) since the late 1970s. The downside of centralised units and functions is also recognised. They are often seen as threatening by other units of administration with attendant resistance to change and potentially strained relationships.

6.6. Organisation of evaluation in public agencies

How evaluation is organised appears to be an important factor in effective evaluation take up and implementation. There are a number of major topics within this debate:

(a) it is a matter of values and attitudes, a belief that it is right to look critically at policies and programmes and to gather and consider evidence about what works and how to improve performance;

(b) it is a matter of administrative practice, how administrations are organised, how stakeholders are involved in the evaluation process and how appropriate levels of separation and integration are maintained between those who implement and those who evaluate public policies;

(c) it is a matter of system integration; systems that are supportive of an evaluation culture are usually networked through professional associations and adhere to common professional standards.

In the section that follows, (which continues to draw mainly on Boyle and Lemaire), we concentrate mainly on the administrative arrangements.

Decentralised evaluation units are most likely to be intended to support decision-making and programme effectiveness at a programme level. They are usually aligned with programme management and are hence less threatening. However, issues of independence and bias can arise from their closeness to programme personnel. Another weakness may be lack of methodological skills in evaluation, though the most important criticism raised by Sonnichsen is the possible lack of power that decentralised units have, especially where decisions about programme and policy futures are still made at a centralised level. Ultimately this debate resolves itself into one of evaluation purpose. Where the purpose is primarily to improve programme and policy implementation, there appear to be strong arguments for a decentralised model which will also favour learning at a decentralised level. Where the purpose is primarily to support central strategies and policy-making, the argument for a centralised model appears to be stronger: here the learning would tend to occur centrally rather than at programme or policy division level. Issues of professional competence and skills acquisition appear to be stronger within a decentralised framework. This issue is further examined in the same volume by Boyle, who discusses the human resources aspect of evaluation specialists and the professionalisation of evaluation. He highlights, in particular, the importance of appropriate training courses and education curricula, the way expertise is deployed mixing different skills together, and the importance of continuous in-service training. From the point of view of this review, much of this discussion is generic rather than focused on use. However, from a use perspective, Boyle makes a strong argument for developing evaluation users. Evaluation users also need to be trained and developed to become consumers. In some ways this can be seen as one of the consequences also of introducing managing for results approaches in public service organisations. This, in Boyle's terms, creates the link between the supply and demand sides of evaluation use.

The shift in the institutional practice of evaluation within the public sector from monitoring and management to accountability and performance is widely noted (see especially Bastoe, 1999) One definite tendency that this and other studies demonstrate is the integration of evaluation and monitoring with various approaches to performance management.

Among the elements of this approach, following an influential OECD paper (OECD, 1995), are:

(a) objective and target setting;
(b) management responsibility to implement against targets;
(c) the monitoring of performance;
(d) the feed-in of such performance into future policy-making and programming;
(e) the provision of information to external parliamentary and audit committees for *ex-post* review.

There is also disagreement about the need and benefit of linking audit and evaluation, which Bastoe also recognises, especially when implementing performance management systems. It is, for example, important also to consider 'how learning actually takes place in organizations'. This attention to learning is also the preoccupation of other authors in the institutionalisation of evaluation in the public sector.

6.7. Evaluation as dialogue

A quite different approach to analysing evaluation use in policy settings is suggested by Van der Knaap (1995) in his article *Policy evaluation and learning: feedback, enlightenment or argumentation?* He challenges the traditional rational-objectivist model of policy evaluation, favouring rather a constructivist view in which policy-makers conduct dialogues about evaluation findings in order to reach their conclusions. Thus 'policy-making is conceived of as an ongoing dialogue, in which both governmental and societal actors contest their views on policy issues by exchanging arguments'. At heart, this argument challenges the 'positivist idea that policy evaluators may provide the policy-maker with neutral or objective feedback information or recommendations'. Rather than enlighten the policy-maker, 'at best, the evaluator might contribute to the quality of policy discourse by entering the negotiations that compose the policy-making processes with informed arguments and a willingness to listen, argue, and persuade or be persuaded'. This shift from the rational to the argumentative is, according to Van der Knaap, a way to 'institutionalise policy orientated learning'. This is not to suggest that the evaluator is relieved of the responsibility to provide reliable information and findings but that there is a need also to supplement traditional analysis with material that will stimulate debate and allow different stakeholders to consider material presented from different perspectives.

A similar logic informs a recent article by Valovirta (2002). 'Rather than regarding evaluative information as indisputable knowledge, it is viewed as a collection of arguments, which can be debated, accepted and disputed'. According to the author, utilisation of evaluation should be regarded as a process that runs through four stages:

(a) familiarisation with evaluation results and involvement in the evaluation process;
(b) interpretations based on expectations, assessments of the quality of research (truth test) and the feasibility of actions proposed or implied (utility test);
(c) argumentation in which 'individual interpretations are […] subject to collective deliberation, discussion, negotiation and decision-making';
(d) effects which may take the form of decisions and actions, new shared comprehensions and increased awareness; and increased or undermined legitimacy.

Within this perspective on evaluation use, 'evaluations force people to present well-grounded arguments for refuting evaluation conclusions and recommendations. This opens up possibilities for new understandings to emerge.' (Valovirta, 2002; p. 77).

Valovirta makes an important distinction between the different contexts in which evaluation takes place. In particular he distinguishes between settings where there is a high level of consensus versus those with a high level of conflict; and settings where there is a low pressure for change versus those with a high pressure for change. He suggests that these contextual differences will determine the nature of the argumentation that takes place around evaluation findings.

6.8. Strategies and types of evaluation use

In summary, and cutting across the theoretical debates described above, we can identify six main strategies or approaches to evaluation use from this discussion:

(a) instrumentalist, when evaluation is used instrumentally to achieve an intended and explicit type of use, e.g. make recommendations that are then implemented;

(b) incrementalist, when evaluation becomes useful cumulatively over time by bringing together evaluation findings from different evaluations e.g. through meta-analysis and synthesis, in order to inform action;

(c) process-oriented, when evaluation is useful as much because of the processes of engagement and debate it engenders among stakeholders as because of the results it produces: thinking changes even if recommendations are not implemented;

(d) administrative proceduralist, when the procedures through which evaluation is organised and delivered make evaluation use more likely, e.g. well structured terms of reference plus requirements that programme managers act on or, at least, respond to evaluation findings;

(e) systemic proceduralists, when the wider system in which evaluation is embedded including dissemination networks, communities of practice and administrative cultures encourage evaluation use;

(f) performance management, when the demands of improving administrative performance and achieving targets creates a market for evaluation outputs.

In the real world, these categories are not mutually exclusive and most public agencies will pursue more than one. However, they do constitute major alternatives; few users of evaluation attempt many of these strategies simultaneously. Some of these different approaches to evaluation use are more or less supportive of different types of evaluation. For example, instrumental approaches are more likely to be consistent with an accountability or outcome and impact type of evaluation; a performance management approach is mainly concerned with improving and developing programmes, evaluation being formative for these programmes. However different contexts of use can support very different types of evaluation; processual strategies may conduct their debates and arguments about evaluations that are neither concerned with processes nor set out to be formative. The main implication from the above discussion is that evaluation is not simply an applied technology or method. Rather it is embedded in contexts of use that shape what evaluation becomes in particular contexts and fields of application.

7. Evaluation standards and regulation

7.1. Evaluation codes and standards

Questions about the roles of evaluators are unavoidable, partly because the dominant logic of evaluation sees them exercise so much control. As with any other debate about the responsibility of professionals (doctors, auditors, research scientists) the question of *quis custodiet ipso custodes?* is soon heard. Given that the usual answer among professionals is 'through self-regulation', the subsequent question is: 'through what means and against what criteria and standards does regulation occur?'. It is in order to establish some shared agreement among evaluators, commissioners and those affected by evaluation, that evaluation codes and standards have become so prevalent. However, codes and standards also have a broader purpose. In a decentralised system composed of many stakeholders, standards are a way of regulating behaviour across organisational boundaries, provided, that is, that all parties accept these norms.

As the previous section has tried to show, there is widespread concern among evaluators about evaluation use. This has partly been fuelled by the academic debates reviewed above, which have highlighted the problem of evaluations not being used. This is one impetus behind the development of guidelines and standards, usually for evaluators but also for commissioners, that are intended to promote evaluation use directly and indirectly. They are direct because they often concern use. They are indirect because they always seek to enhance the quality of evaluation, which is widely assumed to be a factor in enhancing evaluation use and practice more generally.

The earliest of these efforts have been produced by the AEA (1995) and its precursor organisations, for example the Joint Committee on Standards for Educational Evaluation (1994).

The programme *Evaluation standards*, a guide for evaluators, particularly from a programme perspective, identifies standards under four main headings:

(a) utility standards 'are intended to ensure that an evaluation will serve the information needs of intended users';

(b) feasibility standards 'are intended to ensure that an evaluation will be realistic, prudent, diplomatic, and frugal';

(c) propriety standards 'are intended to ensure that an evaluation will be conducted legally, ethically, and with due regard for the welfare of those involved in the evaluation, as well as those affected by its results';

(d) accuracy standards 'are intended to ensure that an evaluation will reveal and convey technically adequate information about the features that determine worth or merit of the program being evaluated' (from Joint Committee on Standards for Educational Evaluation, 1994).

These standards are variously concerned with sound methods, timely dissemination, the independence and impartiality of evaluators and the necessary level of evaluator skill and competence.

Perhaps the clearest indication of how such standards are intended to establish norms that will influence the conduct of evaluation is to be found in the way 'utility' is elaborated as a standard. Thus utility includes:

(a) being clear about stakeholders so that their needs can be addressed;

(b) ensuring the credibility of the evaluators so that their results are likely to be accepted;

(c) collecting relevant information from a broad range of sources as understood by clients and stakeholders;

(d) being clear about value judgement used to interpret findings;

(e) reporting clarity so that the information provided in reports is easily understood;

(f) disseminating reports to intended users in a timely fashion;

(g) planning evaluations from the outset in a way that encourages follow-through from stakeholders.

It is worth noting that other standards also have implications for the conduct of evaluation. For example, feasibility standards include a substandard entitled political viability. This is concerned with obtaining the cooperation of different interest groups in order to limit bias or misapplication of results. Similarly the propriety standards include a substandard entitled service orientation which is designed to assist organisations to address and effectively serve the needs of a full range of targeted participants; and one entitled conflicts of interest which seeks to avoid compromising evaluations and their results. Accuracy standards are also relevant, for example valid information, one of the substandards of accuracy, is justified in terms of assuring the interpretation arrived at is valid for the intended use.

This set of standards has been widely imitated and adapted to different national contexts including, most recently, by the *Deutsche Gesellschaft für Evaluation* (DeGEval; 2002; see also Beywl and Speer in this report). There are also discussions and plans in France and the UK to develop standards. These discussions have, however, moved away from directly adopting the North American model.

Alongside the widespread adoption of the AEA *Programme Evaluation Guidelines* there has also been the emergence more recently of a set of guidelines for the ethical conduct of evaluations. These differ from programme guidelines, being more concerned with the ethical dilemmas that both commissioners and evaluators face in the course of the evaluation process. Such guidelines have been variously prepared by the Australasian Evaluation Society (AES, 1997), the Canadian Evaluation Society (CES) ([1]) and the AEA. These ethical guidelines are relevant to a discussion about evaluation itself because they, too, are concerned with the credibility as well as the feasibility of evaluation.

Both the CES and the AEA direct their attention to evaluators. Thus the CES guidelines, under the three main headings of competence, integrity and accountability, are concerned that evaluators should be competent, act with integrity and be accountable. Similarly the AEA in its *Guiding principles for evaluators* – under various headings of systematic inquiry, competence, integrity/honesty, respect for people, responsibilities for general and public welfare and recommendation for continued work – also direct their attention to what evaluators ought to do.

One set of ethical guidelines drawn up by evaluators' professional bodies, produced by the AES, stand out from the others. These are concerned with ethical behaviour and decision-making among commissioners, users and teachers of evaluation as well as evaluators themselves. According to the AES, the primary groups addressed by their guidelines are commissioners and evaluation teams or evaluators. In this regard also the AES diverges from some of the other guidelines referred to above. They acknowledge that evaluators often work in teams rather than mainly as individuals.

The AES *Guidelines for the ethical conduct of evaluations* follows the evaluation cycle by grouping its guidance under three main headings:

(a) commissioning and preparing for an evaluation;
(b) conducting an evaluation;
(c) reporting the evaluation results.

One view one can take about the AES guidelines is that they constitute a quality assurance framework for the entire evaluation process. These guidelines focus on a number of ways in which credibility of evaluations can be enhanced, i.e.:

(a) shared expectations between evaluators and commissioners about what can be delivered through an evaluation;
(b) strengthening the basis for evaluation judgements;
(c) reducing conflicts during the course of the evaluation;
(d) ensuring balance and simplicity in the way reports are presented.

However, it should be noted that the AES regards these guidelines as complementary to, rather than as a substitute for, other guidelines such as the *Programme evaluation guidelines*; indeed they encourage the use of the *Guidelines for the ethical conduct for evaluations* jointly with the *Programme evaluation standards*.

A set of evaluation guidelines and standards that is firmly set within the concerns of the programme managers and commissioners of evaluation, has recently been issued by the European Commission ([2]). These are linked to the introduction of 'activity based management', the Commission's form of results-based management, and are consistent with many of the approaches to strengthening evaluation capacity and evaluation use discussed above. They cover how evaluation 'functions' across the Commission should be organised and resourced, how evaluations should be planned and managed,

([1]) www.evaluationcanada.ca

([2]) http://europa.eu.int/comm/budget/evaluation/index_en.htm how results should be used and disseminated and how to ensure good quality reports.

There are a variety of other guidelines for evaluation that are indicative of the widespread interest in evaluation standards. For example, the *Means collection* (European Commission, 1999), Volume 1, *Evaluation design and management*, has a section on optimising the use of evaluation. This focuses on dissemination, distinguishing between both different communication channels (e.g. reports, synthesis, article, confidential note) and audiences (e.g. commissioners of the evaluation, steering groups, managers, European institutions, citizens and journalists). The *Means collection* authors also acknowledge the 'absence of a direct short-term link between recommendations and decisions'.

Evaluation standards are not only intended as a framework for the design of particular evaluations. They are also used in meta-evaluations, to try to describe the range of evaluation practice, and as a quality assessment tool in relation to completed evaluations. In the European Commission, for example, the means 'quality criteria' are widely used as a framework for assessing evaluations, both at proposal and completion stages (Table 5 below).

Table 5: **Grid for a synthetic assessment of quality of evaluation work**

With regard to this criterion, the evaluation report is 1. unacceptable 2. acceptable 3. good 4. excellent	1	2	3	4
Meeting needs: does the evaluation adequately address the requests for information formulated by the commissioners and does it correspond to the terms of reference?	□	□	□	□
Relevant scope: have the rationale of the programme, its outputs, results, impacts, interactions with other policies and unexpected effects been carefully studied?	□	□	□	□
Defensible design: is the design of the evaluation appropriate and adequate for obtaining the results (with their limits of validity) needed to answer the main evaluative questions?	□	□	□	□
Reliable data: are the primary and secondary data collected or selected suitable? Are they sufficiently reliable compared to the expected use?	□	□	□	□
Sound analysis: are quantitative and qualitative data analysed in accordance with established rules, and are they complete and appropriate for answering the evaluative questions correctly?	□	□	□	□
Credible results: are the results logical and justified by the analysis of data and by interpretations based on carefully presented explanatory hypotheses?	□	□	□	□
Impartial conclusions: are the conclusions just and non-biased by personal or partisan considerations, and are they detailed enough to be implemented concretely?	□	□	□	□
Clear report: does the report describe the context and goal, as well as the organisation and results of the evaluated programme in such a way that the information provided is easily understood?	□	□	□	□
In view of the contextual constraints bearing on the evaluation, the evaluation report is considered to be	□	□	□	□

Source: European Commission, 1999

On the boundaries of the territory covered here, is a wider literature on innovation and good practice. Thus in many studies that have been conducted at a European level there is a concern for dissemination of innovation. An interesting example of this is found in Mannila et al. (2001). In this volume, Sonderberg discusses mainstreaming, innovation and knowledge networks. The way in which innovations are spread is close to the problem of how to encourage the take-up of the results of evaluations. Sonderberg also focuses on knowledge networks (which is close to notions of policy communities and communities of practice) as a means of ensuring the dissemination and mainstreaming of new ideas. Arguably this broader literature on the diffusion of innovation could be regarded more generally as a source of insights into how evaluation practice and findings are disseminated.

While there has been a great deal of discussion around standards and codes in North America, Australasia and, most recently, in Europe, the evidence of take-up of such standards and codes is sparse. Most evaluators can come up with examples of 'breaches' in most of their evaluations. The systematic application of standards is rare.

7.2. Evaluation as a method for regulating decentralised systems

There is a need to understand the pervasiveness of evaluation across sectors, branches of public administration and the public sector more widely and its growth as a practice. As was suggested in the terms of reference for this study, part of the answer rests with the nature of public sector reform, including deregulation and decentralisation, that has occurred over the last two decades. This process has fragmented previously established mechanisms of control and management. No longer are programmes, let alone policies, delivered by single agencies. The *de facto* norms and standards that characterised well-established public bodies are no longer shared among those responsible for programmes and policy instruments. This is important not only for retrospective accounting for success or failure but also to plan and implement major interagency initiatives. Evaluation has become one of the key elements in the new technology of coordination and interagency management required by the restructured public sector.

This has been noted by evaluation researchers in the education sector in particular. Thus Henkel (1997) has noted how the external assessment of higher education institutions in the UK has led to a reduction in their traditional autonomy and arguably a shift towards increased managerial intervention inside these institutions. Segerholm (2001) develops a more extensive theory of 'national evaluations as governing instruments'. She describes how the evaluation of education programmes in Swedish Higher Education 'worked as an administrative technique for disciplining and […] governing'. In this, Segerholm follows Foucault in arguing that diffused evaluation systems with criteria that are internalised by institutions and practices that make the application of such criteria visible, increase opportunities for central control. Similar arguments have been advanced in other parts of the public sector that have been subject to wholesale reform and decentralisation – but still held accountable – especially in northern European and North America.

A second explanation for the spread and take-up of evaluation has been the new demand for accountability that comes from better informed and less deferential citizenry. Those responsible for delivering public policy are faced by conflicting demands. On the one hand their political masters require success, transparent operations, evidence of efficiency and explanations of failure. On the other hand, communities, presumed beneficiaries and those affected by policy initiatives, schooled in the market rhetoric of customers who receive a service, are increasingly sceptical of government action. The growth in ever more detailed management oriented and explanatory evaluation frameworks, and the simultaneous growth of participative, bottom up, stakeholder led evaluation practice, is a response to these contradictory demands.

Evaluation in these contexts can be seen as a control process within the cybernetic meaning of the term. Terms such as steering, guidance and regulation are frequently attached to descriptions of evaluative activity. In these terms, standards become especially important. Standards become not only a tool for judging and valuing, they also become the means to develop and express consensus among fragmented policy communities and communities of practice.

The role of standards in evaluation, therefore, needs to be seen much more broadly than the debate about the standards that apply to evaluation practice and outputs. Four main understandings of standards in evaluation can be identified:

(a) those concerned with standards (and criteria) to judge outcomes and effectiveness;
(b) those concerned with required standards of performance in decentralised administrative systems engaged in programme delivery;
(c) those concerned with devolved self-evaluating systems operating within a predetermined framework;
(d) those concerned with the required behaviour of evaluators and those who commission evaluation.

Each of these has been discussed in this chapter. However, the nature of these standards differ subtly, for example:

(a) evaluation standards allow for judgements to be made based on norms and/or the beliefs of stakeholders;
(b) programme delivery standards are usually set by higher levels of an administration (or possibly through regulation) in order to exercise influence over the performance of others;
(c) the standards that operate in devolved systems are processual; they concern the obligation of those to whom powers are devolved to follow certain procedures, including making evaluation outputs available;
(d) standards for evaluators are part of the self-regulation agenda or a recently emerged professional group, mainly developed from within that profession.

Types of evaluation identified in this chapter in part derive from research and conceptualisation about evaluation. However this research-based activity interacts with two others. First, evaluators as practitioners reflect on their own practice and have become more professional to the point where issues of professional self-regulation have been highlighted. Second evaluation is shaped by a complex web of contractual and institutional demands. Both of these activities, evaluators as reflective practitioners and evaluation as an institutionalised and market/network based practice, determine what types of evaluation survive as well as what types are possible or advocated.

8. Conclusions and future scenarios

As this study is conceived within a European context, there are also questions to be asked as to the likely evolution of evaluation within Europe. This is not to suggest that evaluation in Europe is *sui generis*, rather that there are distinctive aspects of the EU's institutions and the challenges that evaluators face in Europe. This distinctiveness is shaped by a number of underlying dynamics, the most important of which are discussed below.

The first is accountability and participation. The policy system, the main customer of evaluation, faces contradictory demands. On the one hand, the demand for accountability for monies spent and for promises made generates mounting pressure for top-down, performance oriented, quasi-audit types of evaluation. On the other hand, citizens who often feel estranged from their politicians and question the basis of distant, context-free decisions, demand closer involvement and participation in the way policies are designed and implemented. The extent to which evaluation evolves in an accountability or participative direction will depend on how these contradictory demands are – or are not – resolved. This is not to suggest that both variants cannot coexist, but the two cultures of evaluation are very different and the balance of thinking and resource constitute genuinely alternative scenarios.

Diversity and convergence is a second consideration. The ever-expanding European 'space' is increasingly diverse in its traditions, institutions, languages and culture. The European project was founded on a vision of integration that assumes convergence. Even though the contrary vision of subsidiarity emphasises continued diversity as much as convergence, both tendencies remain strong. This matters for evaluation because the diversity/convergence dynamic underpins many of the roles that evaluators are expected to take-up. In particular it affects how evaluators relate to stakeholders. When diversity is accepted, value-sensitivity and working closely with stakeholders is taken for granted. When convergence is assumed there is more likely to be a presumption of homogeneity among those affected by public intervention and a tendency to favour methods and *modus operandi* that take limited account of the particular circumstances of policy and programme implementation.

Reconciliation of evaluation cultures is also important. Alongside the cultural diversity of Europe, and not disconnected from it, is cultural diversity among evaluators. Within the social and economic sciences in particular, there are familiar cleavages, many of which have been discussed in this chapter. For example, evaluation also has its advocates for largely empirical and positivist methodologies as well as advocates for theory-based investigation in various forms. What constitutes evidence, validity, generalisability and legitimate conclusions are hotly debated among evaluators as they are in other academic and professional circles. To an extent, these cultural divides among evaluators mirror, and are reinforced by, cultural divisions within Europe. The philosophical patrimony of Latin, Scandinavian and Anglo-Saxon countries, can make them receptive to different evaluation approaches and methodologies. There is also a history of policy borrowing within Europe, with established patterns of shared professional and policy networks that predispose some countries to adopt practices more easily from particular countries rather than others. (Italians are more likely to 'borrow' from the French, Scandinavians from each other and the British from North Americans.) So, the alternatives of convergence or national – or at least regional – specificity are available to evaluators as they are to Europe's citizens and their Member States. Among the influences that will shape these alternatives, the growth of evaluation societies at a European level and the policies of the EU, are likely to be important, as are the dissemination of evaluation standards and procedures by European institutions.

Solidarity and social exclusion also feature. In many socioeconomic programmes, including VET, lack of social solidarity is a factor that both underpins the problems that the programme seeks to address and constrains the policy responses that are possible. *In extremis* social exclusion is at the heart of these programmes. However, the methods of evaluation that are frequently deployed assume solidarity and make little contribution towards social cohesion. From within a presumption of solidarity, evaluators can easily become part of a regulatory or control regime. A counter tendency within contemporary evaluation is to engage actively in formative, developmental and trust building activities as part of the evaluation itself. This goes beyond participation or offering stakeholders a voice: it takes the opportunity to support social inclusion and solidarity by the way an evaluation is conducted. For example, not only do evaluators study the realities of partnership working, they also contribute to the strengthening of partnerships by the methods and research strategies they adopt;

There is also tension between complexity and linearity. Characteristically, much evaluation work is increasingly complex. Such complexity underpins a high level of uncertainty in policy interventions. Success is unsure and goals need to be redefined (or rationalised) along the way. However, many policy interventions are simple in their logic. They assume a linear input leading to a predictable output: increased investment in VET leads to higher wages, or more work-based learning leads to greater competitiveness. When reinforced by a strong accountability ethos, this places pressure on evaluators to attempt to calibrate the future (e.g. through impact assessment) and demonstrate through ‘success stories’ early wins and the achievement of anticipated goals. When the policy system is able to acknowledge uncertainty – which implies an unusual degree of openness to citizens and electorates – new possibilities are opened up also for evaluators. They become part of a more reflexive and iterative ‘learning culture’, with lessons learned from mistakes, as well as from evidence of success.

Finally, there continues to be important divisions between policy-makers and evaluators. Some of these derive from the dynamics identified above. For example, the ethos of accountability and the need to demonstrate success, which drives policy-makers, and the preoccupation with complexity and value difference, which tends to drive evaluators to a greater extent. Bridging the cultural divide between policy-makers and evaluators has become a theme in many European debates about evaluation policy and practice. In the first generation of these debates, the emphasis has been on educating evaluators to understand the priorities and pressures on policy-makers: the importance of deadlines, the need for clear recommendations, the need to accept the parameters of current conventional wisdom. This is often accompanied by a demand from evaluators that ‘policy-makers become more like us’. The emergence of networks and communities of practice that span both evaluators and policy-makers opens up a second generation of debate. This offers the possibility of shared frameworks, an understanding of what evaluation can and cannot achieve and an understanding of what policy-makers need in order to learn from and to use evaluation outputs and processes. The extent to which this second generation debate becomes more commonplace, so that European evaluators and policy-makers can move beyond the ‘why can’t they be more like us’ refrain, will also shape the way in which European evaluation evolves in the future.

Considering the possible futures of evaluation brings into focus many of the issues raised elsewhere in this article, but from a particular perspective. While evaluators, policy-makers and evaluation researchers advocate different evaluation approaches, models and practices, what this concluding section has sought to emphasise is that evaluation itself is shaped by societal dynamics and contingencies. There are choices between: types of evaluation and underlying philosophies of science; capacity development and evaluation use; evaluation theory and theory more generally; and the prominence of participation, empowerment and self-regulation on the one hand, and top-down, policy driven variants of evaluation on the other. These are not open and unencumbered. Rather like other (social) technologies, evaluation is shaped by wider societal, political and institutional dynamics. The future of evaluation in Europe, as elsewhere, should therefore be understood also in terms of the bigger picture: of the way policy systems adapt to socioeconomic, cultural and political challenges, and the way evaluators themselves engage as actors, making choices about how they can contribute to these challenges.

List of abbreviations

AEA	American evaluation association
AES	Australasian evaluation society
CES	Canadian evaluation society
CMO	Context, mechanism, outcome
CVT	Continuing vocational training
RCTs	randomised control trials
VET	Vocational education and training

References

Alkin, M. C. *Debates on evaluation*. London: Sage, 1990.

AEA – American Evaluation Association Task Force on Guiding Principles for Evaluators. Guiding Principles for Evaluators. In: Shadish, W. R. et al. (eds) *New directions for program evaluation – guiding principles for evaluators*, 1995, No 66, p. 19–24.

AES – Australasian Evaluation Society. *Guidelines for the ethical conduct of evaluation*, 1997. Available from Internet: http://www.aes.asn.au/aesguide02.pdf [Cited 30.6.2003].

Biott, C.; Cook, T. Local evaluation in national early years excellence centres pilot programme. *Evaluation – The International Journal of Theory, Research and Practice*, 2000, Vol. 6, No 4, p. 399-413.

Boyd, R. Confirmation, semantics and the interpretation of scientific theories. In: Boyd, R.; Gasper, P.; Trout, J. D. *The philosophy of science*. Cambridge, MA: MIT Press, 1991.

Bastoe, P. O. Linking evaluation with strategic planning, budgeting, monitoring and auditing. In: Boyle, R.; Lemaire, D. (eds) *Building effective evaluation capacity: lessons from practice*. London: Transaction Publishers, 1999, p. 93-110 (Comparative policy analysis series).

Boyle, R.; Lemaire, D. (eds) *Building effective evaluation capacity: lessons from practice*. London: Transaction Publishers, 1999 (Comparative policy analysis series).

Caracelli, V. J.; Preskill, H. (eds) *New directions for evaluation: the expanding scope of evaluation use*. San Francisco: Jossey-Bass, 2000.

Chelimsky, E. Politics, policy and research synthesis. *Evaluation – The International Journal of Theory, Research and Practice*, 1995, Vol. 1, No 1 p. 97-104.

Chelimsky, E. Thoughts for a new evaluation society. *Evaluation – The International Journal of Theory, Research and Practice*, 1997, Vol. 3, No 1, p. 97-118.

Chen, H.-T. *Theory-driven evaluations*. London: Sage, 1990.

Connell, J. P. et al. (eds) *New approaches to evaluating community initiatives: concepts, methods and contexts*. New York: The Aspen Institute, 1995.

Cronbach, L. J. *Designing evaluations of educational and social programs*. San Francisco: Jossey-Bass, 1982.

Cronbach, L. J. Construct validation after thirty years. In: Linn, R. L. (ed.) *Intelligence: measurement, theory and public policy*. Urbana: University of Illinois Press, 1989.

DeGEval – Deutsche Gesellschaft für Evaluation [German Evaluation Society]. *Standards für Evaluation*. Köln: DeGEval, 2002 Available from Internet: www.degeval.de/standards/index.htm [cited 30.6.2003].

Davies, H. T. O.; Nutley, S. M.; Smith, P. C. (eds) *What works?* Bristol: The Policy Press, 2000.

European Commission. Evaluation design and management. In: *Means collection: evaluating socio-economic programmes*. Luxembourg: Office for Official Publications of the European Union, 1999, Vol. 1.

Fetterman, D. M.; Kaftarian, S. J.; Wandersman, A. (eds) *Empowerment evaluation: knowledge and tools for self-assessment and accountability*. London: Sage, 1996.

Fournier, D. M. Establishing evaluative conclusions: a distinction of general and working logic In: Fournier, D. M. (ed.) *Reasoning in evaluation: inferential links and leaps*. San Francisco: Jossey-Bass, 1995, p. 15–32 (New directions for program evaluation, 68).

Fulbright-Anderson, K.; Kubisch, A. C.; Connel, J. P. *New approaches to evaluating community initiatives*, Vol. 2: Theory, measurement and analysis. Washington DC: The Aspen Institute, 1998.

Gallo, P. S. Meta-analysis – a mixed metaphor. *American Psychologist*, 1978, Vol. 33, p. 515–517.

Guba, E. G.; Lincoln, Y. S. *Fourth generation evaluation*. London: Sage, 1989.

Henkel, M. Teaching quality assessments: public accountability and academic autonomy in higher education. *Evaluation*, 1997, Vol. 3, No 1, p. 9–23.

House, E. R. Assumptions underlying evaluation models. In: Madaus, G. F.; Scriven, M.; Stufflebeam, D. L. *Evaluation models viewpoints on educational and human services evaluation*. The Hague: Kluwer-Nijhoff Publishing, 1983.

Joint Committee on Standards for Educational Evaluation. *The program evaluation standards. How to assess evaluations of educational programs*. Thousand Oaks: Sage, 1994.

Julnes, G.; Mark, M. M.; Henry, G. T. Promoting realism in evaluation: realistic evaluation and the broader context. *Evaluation – the International Journal of Theory, Research and Practice*, 1998, Vol. 4, No 4, p. 483-504.

Lave, J.; Wenger, E. R. *Situated learning: legitimate peripheral participation*. Cambridge: Cambridge University Press, 1991.

Leeuw, F. L.; Rist, R. C.; Sonnichsen, R. C. *Can governments learn? Comparative perspectives on evaluation and organization learning*. Somerset NJ: Transaction Publishers, 1994 (Comparative Policy Analysis Series).

Leeuw, F. L.; Toulemonde, J.; Brouwers, A. Evaluation activities in Europe: a quick scan of the market in 1998. *Evaluation – the International Journal of Theory*, Research and Practice, 1999, Vol. 5, No 4, p. 487–496.

Mannila, S. M.; Ala-Kauhaluoma, M.; Valjakka, S. (eds) *Good practice in finding good practice: international workshop evaluation*. Naperville, IL: Rehabilitation Foundation, 2001.

Miller, R. W. Fact and method in the social sciences. In: Boyd, R.; Gasper, P.; Trout, J.D. *The philosophy of science*. Cambridge, MA: MIT Press, 1991.

Nagarajan, N.; Vanheukelen, M. *Evaluating EU expenditure programmes: a guide to intermediate and ex-post evaluation*. Luxembourg: Office for Official Publications of the European Communities, 1997.

Nutley, S. M.; Walter, I.; Davies, H. T. O. From knowing to doing: a framework for understanding the evidence-into-practice agenda. *Evaluation – the International Journal of Theory, Research and Practice*, 2003, Vol. 9, No 2.

OECD – Organisation for Economic Cooperation and Development. *Budgeting for results: perspectives on public expenditure management*. Paris: OECD, 1995.

Patton, M. Q. *Utilization-focussed evaluation: the new century text* (3rd Edition). London: Sage, 1997.

Patton, M. Q. *Qualitative research and evaluation methods*. London: Sage, 2002.

Pawson, R. Evidence based policy: in search of a method. *Evaluation – the International Journal of Theory, Research and Practice*, 2002(a), Vol. 8, No 2, p. 157-181.

Pawson, R. Evidence based policy: the promise of 'realist synthesis'. *Evaluation – the International Journal of Theory, Research and Practice*, 2002(b), Vol. 8, No 3, p. 340-358.

Pawson, R.; Tilley, N. *Realistic evaluation*. London: Sage, 1997.

Rist, R. C.; Furubo, J. E.; Sandahl, R.(eds) *The evaluation atlas*. London, 2001.

Rogers, E. *Diffusion of innovation*. New York: The Free Press, 1995.

Rogers, P. J. Program theory: not whether programs work but how they work. In: Stufflebeam, D. L.; Madaus, G. F.; Kellaghan, T. *Evaluation models viewpoints on educational and human services evaluation* (2nd edition). Dordrecht: Kluwer Academic Publishers, 2000.

Rossi, P. H.; Freeman, H. E.; Lipsey, M. W. *Evaluation: a systematic approach* (6th edition). London: Sage, 1999.

Sabatier, P. A. *An advocacy coalition framework of policy change and the role of policy-oriented learning therein*. Policy Sciences, 1988, Vol. 21, p. 129-168.

Sabatier, P. A.; Jenkins-Smith, H. C. *Policy change and learning – an advocacy coalition approach*. Boulder, CO: Westview Press, 1993.

Sanderson, I. Evaluation in complex policy systems. *Evaluation – the International Journal of Theory, Research and Practice*, 2000, Vol. 6, No 4, p. 433-454.

Schwandt, T. Evaluation as practical hermeneutics. *Evaluation – the International Journal of Theory, Research and Practice*, 1997, Vol. 3, No 1, p. 69-83.

Scriven, M. *Evaluation thesaurus* (4th edition). London: Sage, 1991.

Segerholm, C. National evaluations as governing instruments: how do they govern? *Evaluation,* 2001, Vol. 7, No 4, p. 427-438.

Shadish, W. R.; Cook, T. D.; Leviton, L. C. *Foundations of program evaluation: theories of practice*. London: Sage, 1991.

Shadish, W. R.; Cook, T. D.; Campbell, D. T. *Experimental and quasi-experimental designs for generalised causal inference*. Boston: Houghton Mifflin Company, 2002.

Stake, R. *The art of case study research*. London: Sage, 1995.

Stake, R. For all program evaluations there's a criterion problem. *Evaluation – the International Journal of Theory, Research and Practice*, 1996, Vol. 2, No 1, p. 99-103.

Stern, E. The characteristics of programmes and their evaluation. In: OECD. *National programmes in support of local initiatives*. Paris, OECD, 1992.

Stern, E.; Kelleher, J.; Cullen, J. *Preliminary common evaluation framework*. London: The Tavistock Institute, 1992. (Articulate Deliverable, No 1).

Stufflebeam, D. L. Foundational models for 21st century program evaluation. In: Stufflebeam, D. L.; Madaus, G. F.; Kellaghan, T. *Evaluation models viewpoints on educational and human services evaluation* (2nd edition). Dordrecht: Kluwer Academic Publishers, 2000(a).

Stufflebeam, D. L. Professional standards and principles for evaluations. In: Stufflebeam, D. L.; Madaus, G. F.; Kellaghan, T. *Evaluation models viewpoints on educational and human services evaluation* (2nd edition). Dordrecht: Kluwer Academic Publishers, 2000(b).

Toulemonde, J. *Evaluation identities*, 2001.

Valovirta, V. Evaluation utilization as argumentation. *Evaluation – the International Journal of Theory, Research and Practice*, 2002, Vol. 8, No 1, p. 60-80.

Van der Knaap, P. Policy evaluation and learning: feedback, enlightenment or argumentation. *Evaluation – the International Journal of Theory, Research and Practice*, 1995, Vol. 1, No 2, p. 189-216.

Vedung, V. *Public policy and programme evaluation*. New Brunswick, NJ: Transaction, 1997.

W. K. Kellogg Foundation. *Logic model development guide*. Michigan: W. K. Kellogg Foundation, 2000.

Weiss, C. H. *Evaluation research: methods of assessing program effectiveness*. New Jersey: Prentice-Hall, 1972.

Weiss, C. H.; Bucuvalas, M. J. *Social science research and decision-making*. New York: Columbia University Press, 1980.

Weiss, C. H. Nothing as practical as good theory: exploring theory-based evaluation for comprehensive community initiatives for children and families. In: Connell, J. P. et al. (eds) *New approaches to evaluating community initiatives: concepts, methods and contexts*. New York: The Aspen Institute, 1995.

Weiss, C. H. The interface between evaluation and public policy. *Evaluation – the International Journal of Theory, Research and Practice*, 1999, Vol. 5, No 4, p. 468-486.

Weiss, C. H.. Theory-based evaluation: theories of change for poverty reduction programs. In: Feinstein, O; Picciotto, R. (eds) *Evaluation and poverty reduction*. Washington: World Bank, 2000.

Weiss, C. H. Using research in the policy process: potential and constraints. *Policy Studies Journal*, 1976, Vol. 4, p. 224-228.

Wholey, J. S. Using evaluation to improve program performance. In: Levine, R. A. et al. (eds) *Evaluation research and practice: comparative and international perspectives*. Beverly Hills, CA: Sage, 1981.