

EN

The foundations of evaluation and impact research

Third report on vocational
training research in Europe
Background report



The foundations of evaluation and impact research

Third report on vocational training research in Europe: background report

Pascaline Descy,
Manfred Tessaring (eds)

A great deal of additional information on the European Union is available on the Internet.
It can be accessed through the Europa server (<http://europa.eu.int>).

Cataloguing data can be found at the end of this publication.

Luxembourg: Office for Official Publications of the European Communities, 2004

ISBN 92-896-0278-3

© European Centre for the Development of Vocational Training, 2004
All rights reserved.

Printed in Italy

The **European Centre for the Development of Vocational Training** (Cedefop) is the European Union's reference centre for vocational education and training.

We provide information on and analyses of vocational education and training systems, policies, research and practice.

Cedefop was established in 1975 by Council Regulation (EEC) No 337/75.

Europe 123
GR-57001 Thessaloniki (Pylea)

Postal Address:
PO Box 22427
GR-55102 Thessaloniki

Tel. (30) 23 10 49 01 11
Fax (30) 23 10 49 00 20
E-mail: info@cedefop.eu.int
Homepage: www.cedefop.eu.int
Interactive website: www.trainingvillage.gr

Edited by:

Cedefop

Pascaline Descy, Manfred Tessaring, Project managers

Published under the responsibility of:

Johan van Rens, Director

Stavros Stavrou, Deputy Director

Content

Preface	7
Philosophies and types of evaluation research <i>Elliot Stern</i>	9
Developing standards to evaluate vocational education and training programmes <i>Wolfgang Beywl, Sandra Speer</i>	51
Methods and limitations of evaluation and impact research <i>Reinhard Hujer, Marco Caliendo, Dubravko Radić</i>	131
From project to policy evaluation in vocational education and training – possible concepts and tools. Evidence from countries in transition <i>Evelyn Viertel, Søren P. Nielsen, David L. Parkes, Søren Poulsen</i>	191
Look, listen and learn: an international evaluation of adult learning <i>Beatriz Pont, Patrick Werquin</i>	237
Measurement and evaluation of competence <i>Gerald A. Straka</i>	263
An overarching conceptual framework for assessing key competences. Lessons from an interdisciplinary and policy-oriented approach <i>Dominique Simone Rychen</i>	313
List of authors	332

Other volumes of the background report

The background report to the third Research report is composed of two other volumes published separately the content of which is detailed below.

The foundations of evaluation and impact research

Descy P., Tessaring M. (eds)

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, 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

Evaluation of systems and programmes

Descy P., Tessaring M. (eds)

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

Uwe Lauterbach, Hermann-Günter Hesse, Wolfgang Hellwig, 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

Preface

The series of reports on vocational education and training (VET) research have been published by Cedefop since 1998 ⁽¹⁾. The reports provide a comprehensive review of current research in initial and continuing VET in Europe, research results and their implications for policy, practice and future research. Attention is also paid to the theoretical and methodological foundations and due reference given to relations with institutional, economic, social, demographic, and other fields of social action.

Each research report consists of a background report of several volumes with contributions from renowned researchers (this publication) and a synthesis report elaborated by Cedefop experts.

Third research report: evaluation and impact of education and training

This third report informs on current research on evaluation and the impact of education and training on individuals, enterprises and society and economy in general, including:

- (a) the assessment of education and training systems;
- (b) the implementation and outcomes of programmes and reforms with a VET component;
- (c) impact and cost-benefit research, quantifying the contribution of education, training and skills on, for example, earnings, economic growth, employment and social inclusion.

The report serves both to inform and improve policy and practice, and further develop research. It also contributes to the discussion on the overall European goals expressed by the European Council at its Lisbon and follow-up summits.

Demonstrating the contribution of education and training towards realising a knowledge-based society and specifying the diverse benefits at all levels as done by 'summative evaluation' or impact research – is as important as indicating

ways to improve the design and implementation of education and training programmes or measures by 'formative evaluation'.

The background report

The background report gathers contributions from renowned experts and researchers. They allow the reader to approach evaluation and impact research from various angles: individual, enterprise and macro-system level by also considering essential basics on the foundation, approaches, standards – and limitations – of evaluation and impact research.

Contributions have been regrouped into three broad themes, published in separate volumes:

- (a) impact of education and training;
- (b) the foundations of evaluation and impact research;
- (c) evaluation of systems and programmes.

The foundations of evaluation and impact research

The present volume addresses the philosophical roots, types and standards as well as methods of evaluation and impact research.

E. Stern discusses in-depth the philosophies underlying evaluation and presents different types of evaluation as well as the basic approaches. He addresses also the question of evaluation standards, which is then detailed by W. Beywl who discusses ethical and normative standards for evaluation practices developed in several countries. He aims at providing an input for the current discussion of evaluation standards at European level.

M. Caliendo and D. Radic present an overview review on the most important evaluation methods and techniques at micro and macro level as well as their limitations. They discuss the pre-requirements, usability and adequacy of various evaluation techniques for different purposes, including data requirements.

⁽¹⁾ Tessaring, 1998; Cedefop, 1998; Descy and Tessaring, 2001a and b.

A set of tools for evaluating VET systems and reforms with particular focus on countries in transition is presented and discussed by E. Viertel et al. They provide an important input for the developing of systemic evaluation approaches, useful not only in Central and Eastern Europe countries but also for a number of several Western Europe countries where evaluation cultures are not yet distinct.

For evaluating externally education and training systems, the thematic reviews organised by the OECD can be considered as a best practice. B. Pontz and P. Werquin discuss its methodology and its impact in the context of the review on adult learning.

G. Straka, at a more micro level, discusses competence measurement and evaluation, while designing and selecting key competences in an international context is addressed by D. Rychen.

**Stavros
Stavrou**

**Manfred
Tessaring**

**Pascaline
Descy**

References

- Cedefop. *Vocational education and training – the European research field: background report* (2 volumes). Luxembourg: Office for Official Publications of the European Communities, 1998 (Cedefop Reference series).
- Descy, P.; Tessaring, M. *Training and learning for competence: second report on vocational training research in Europe: synthesis report*. Luxembourg: Office for Official Publications of the European Communities, 2001a (Cedefop Reference series, 6).
- Descy, P.; Tessaring, M. (eds) *Training in Europe: second report on vocational training research in Europe 2000: background report* (3 volumes). Luxembourg: Office for Official Publications of the European Communities, 2001b (Cedefop Reference series, 3008).
- Tessaring, M. *Training for a changing society: a report on current vocational education and training research in Europe* (second edition). Luxembourg: Office for Official Publications of the European Communities, 1999 (Cedefop Reference series).



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

1. Introduction	12
1.1. Scope of this chapter	12
2. Can evaluation be defined?	13
2.1. Assessing or explaining outcomes	13
2.2. Evaluation, change and values	14
2.3. Quantitative and qualitative methods	15
2.4. Evaluation types	15
2.5. A focus on results and outcomes	17
2.6. Participatory methods and devolved evaluations	18
2.7. Theory and practice	19
3. The object of VET evaluation	21
3.1. The domain of VET	21
3.2. Overarching characteristics of evaluation objects	22
4. Philosophical foundations	24
4.1. Positivism, observation and theory	24
4.2. Scientific realism	24
4.3. Constructivists	26
5. Evaluation theory	30
5.1. Programme theory	30
5.2. Theory based evaluation and theories of change	31
5.3. A wider theoretical frame	32
6. Determinants of evaluation use	34
6.1. Instrumental versus cumulative use	34
6.2. Process use of evaluation	34
6.3. The importance of methodology	35
6.4. Evaluation and learning	35
6.5. The institutionalisation of evaluation	36
6.6. Organisation of evaluation in public agencies	37
6.7. Evaluation as dialogue	38
6.8. Strategies and types of evaluation use	39
7. Evaluation standards and regulation	40
7.1. Evaluation codes and standards	40
7.2. Evaluation as a method for regulating decentralised systems	43
8. Conclusions and future scenarios	45
List of abbreviations	47
References	48

List of tables

Tables

Table 1: Overlaps between methodology and purpose	16
Table 2: Evaluation types	18
Table 3: Rules guiding realistic enquiry and method	26
Table 4: Comparing constructivist and 'conventional' evaluation	27
Table 5: Grid for a synthetic assessment of quality of evaluation work	42

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 inte-

gration 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 Institute (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 voca-

tional 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 Comte, 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 powers
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 socio-economic 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 emphasises 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 expecta-

tion 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 recep-

tive 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 appro-

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 evalua-*

tion 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 judgments;
- (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 evalua-

tors 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.

Developing standards to evaluate vocational education and training programmes

Wolfgang Beywl, Sandra Speer

Abstract

There have been numerous attempts in evaluation research to develop guidelines and standards. The best known are the US standards for program evaluation, established by the Joint Committee on Standards for Educational Evaluations (JC). These standards originated in the school and university sector. Some illustrative examples relate explicitly to the area of initial and continuing vocational training. The goal of the study was to assess the transferability of US standards or the derivative standards of the German Evaluation Society (DeGEval, 2002) to vocational education and training (VET). The study considers the following initial questions:

- (a) does the terminology of the standards match the concepts of European initial and continuing vocational training?*
- (b) are any standards not applicable to initial and continuing vocational training?*
- (c) do European evaluation experts understand and accept the key concepts conveyed (e.g. definition of 'evaluation', differentiation between 'formative' and 'summative' evaluation, purpose of evaluation, etc.)?*
- (d) are there specific national differences which should be considered when adapting the groups of standards?*

The standards of the DeGEval (2002) were chosen as a reference point for the following analysis. Other relevant standards were presented, and reflections on intercultural transferability and applicability to the subject of VET were made. First, VET experts were consulted during further discussions in Germany and Austria. Nobody expressed reservations about the transferability of the standards to VET, and no one proposed adaptation. Second, evaluation experts in widely divergent European countries were sent a questionnaire. The majority of those surveyed have a positive attitude to standards and endorse maximum standards. Pluralistic evaluation appears to be an important quality criterion. The single DeGEval standards are also debated and subject to comment on the basis of criteria found in recent European literature on VET evaluations.

Table of contents

1. Introduction	55
2. General evaluation standards	58
2.1. Background and purpose of evaluation standards	58
2.2. Philosophy of evaluation standards	60
2.2.1. Excursus on the meaning of the word 'standard': minimum vs. maximum standards	61
2.3. Standards for evaluations and guiding principles for evaluators	63
2.4. Evaluation standards and models	64
3. Transferability of standards	67
3.1. Intercultural transferability of standards	67
3.2. Current approaches to development of evaluation standards in Europe	70
3.3. Transferability of evaluation standards to VET	73
4. Dialogue with German and Austrian VET experts on evaluation standards	76
4.1. Focused events with the German Federal Institute for Vocational Training (BIBB)	76
4.2. Focused meeting with the Austrian Federal Institute for Adult Education (BifEb)	78
4.3. Conclusions from the dialogues	79
5. E-mail survey of evaluation experts in Europe	82
5.1. Profession and nationality of respondents	82
5.2. Assessment of existing evaluation standards	84
5.3. Further development of evaluation standards	87
5.4. Summary of survey findings	89
6. Reflections on VET evaluation standards literature	91
6.1. Commentary on the utility standards	92
6.1.1. N1/U1: stakeholder identification	93
6.1.2. N2/U2: clarification of the purposes of the evaluation	94
6.1.3. N3/U3: evaluator credibility and competence	94
6.1.4. N4/U4: information scope and selection	95
6.1.5. N5/U5: transparency of values	96
6.1.6. N6/U6 – report comprehensiveness and clarity: evaluation reports should provide all relevant information and be easily comprehensible	96
6.1.7. N7/U7: evaluation timeliness	97
6.1.8. N8/U8: evaluation utilisation and use	98
6.2. Commentary on the feasibility standards	99
6.2.1. D1/F1: appropriate procedures	99
6.2.2. D2/F2: diplomatic conduct	100
6.2.3. D3/F3: evaluation efficiency	100
6.3. Commentary on the propriety standards	101
6.3.1. F1/P1: formal agreement	101
6.3.2. F2/P2: protection of individual rights	102
6.3.3. F3/P3: complete and fair investigation	103
6.3.4. F4/P4: unbiased conduct and reporting	103
6.3.5. F5/P5: disclosure of findings	103

6.4. Commentary on the accuracy standards	104
6.4.1. G1/A1: description of the evaluand	105
6.4.2. G2/A2: context analysis	105
6.4.3. G3/A3: described purposes and procedures	106
6.4.4. G4/A4: disclosure of information sources	106
6.4.5. G5/A5: valid and reliable information	107
6.4.6. G6/A6: systematic data review	108
6.4.7. G7/A7: analysis of qualitative and quantitative information	108
6.4.8. G8/A8: justified conclusions	109
6.4.9. G9/A9: meta-evaluation	109
6.5. Proposals for expanding existing standards	110
6.5.1. Selection of the evaluation model	110
6.5.2. Selection of suitable methods	110
6.5.3. Explicit reference to evaluation of training programmes	111
7. Summary and outlook	112
7.1. Objectives, questions and method of the study	112
7.2. Results and conclusions	112
7.2.1. Standards for programme evaluation	112
7.2.2. Transferability of standards	112
7.2.3. Results from group discussion on the applicability of DeGEval standards to vocational training	113
7.2.4. Survey of evaluation experts in Europe	113
7.2.5. Reflections on VET evaluation standards literature	114
7.3. Outlook	114
List of abbreviations	116
Annex 1: transformation table	117
Annex 2: questionnaire	118
Annex 3: list of experts answering the e-mail survey	123
References	124

List of tables and figures

Tables

Table 1: Main tasks in an evaluation	60
Table 2: Exemplary models of evaluation by value interpretation	66
Table 3: Particularly culturally sensitive DeGEval standards	69
Table 4: VET examples in the JC standards (1994/2000)	73
Table 5: Primary position in evaluation	83
Table 6: Respondents' professional background	83
Table 7: Respondents' relation to VET	83
Table 8: General assessment of evaluation standards	86
Table 9: Preferred type of standards (minimum vs. maximum)	88

Figures

Figure 1: Subject of this paper	57
Figure 2: Evolution of DeGEval standards	59
Figure 3: Respondents' identification with national professional cultures	84
Figure 4: Respondents' familiarity with various sets of evaluation guidelines	85
Figure 5: Seven points which make evaluations useful	92
Figure 6: Two points which make evaluations feasible	99
Figure 7: Five guidelines which keep evaluations on a straight course	102
Figure 8: Nine components which make evaluations accurate	105

1. Introduction

The market for evaluations in Europe is growing rapidly. More evaluations are being performed, and they are playing a decisive role in shaping policy, particularly government policy.

After many decades of evaluation, much of which was in the area of vocational education and training (VET), a wide spectrum of evaluation models has emerged. Evolution in VET will continue to change evaluation requirements.

‘In a time of deregulation and decentralisation, evaluation becomes increasingly important as a steering mechanism. This makes it vulnerable to misuse. Evaluations can be used as a spurious justification for practices that are deemed politically expedient rather than objectively serving their purpose. This demands a rigorous discipline as well as ethical standards [...]’ (Cedefop, 2001, p. 6).

Nevertheless, no standards are yet recognised as quality requirements and guidelines for evaluations of VET in European Union Member States. After years of experience in evaluation, there continues to be reflection on, and systemisation of requirements for, good evaluation of VET. This desideratum can form a basis for expert discussion.

In this paper, evaluation is broadly defined as ‘systematic investigation of the applicability or merit of an evaluand’ (JC, 2000, p. 25). The uniqueness of the evaluation derives from the fact that concepts, structures, processes and results of programmes are described and graded according to their relationship to target groups or in social systems on the basis of empirical, scientific methods. Evaluation also provides the foundation for impact-oriented programme control.

This paper focuses on the theory and practice of evaluations which address initial and continuing vocational training programmes (¹).

The term ‘programme’ can have various meanings, depending on level of reference, field of study and policy area. A macro-programme, for

instance, can encompass major bundles of VET measures as part of EU policy. By the same token, local continuing training measures and initial training initiatives in individual corporate divisions can become a programme for evaluation. For people-oriented service programmes, which is what most VET measures are, the intended impact only appears in the desired quantity and quality if target group members actively participate (coproduction, *uno actu* principle, Haller, 1998). Systematic ratings and descriptions of human service programmes with a claim to intersubjective reliability are highly vulnerable, given varying, even contrary, economic and social interests and values. This applies to all phases of the evaluation: selection of evaluators, definition of information scope, interpretation of data, drafting of the evaluation report and formulation of conclusions and, in some cases, recommendations.

High-quality evaluations are required to achieve acceptance and credibility of evaluations among programme participants and evaluation report addressees. To do this, and thus to increase the acceptability and utilisation of evaluations, norms, rules, guidelines and standards are devised for evaluations. Nussli (1999, p. 283) writes: ‘A key prerequisite is that education and training evaluation research, which has so far concentrated on scholastic education, should be more involved in the construction of evaluation methods and the development of quality standards and meta-evaluation procedures.’ As a rule, evaluation standards and guiding principles for evaluators spell out organisational, legal, technical and methodological evaluation requirements as well as ethical principles and considerations.

The term ‘standard’ has attracted increased attention recently in education and training circles, not only with reference to evaluation. The Euro-

(¹) ‘Programme’ is a generic term from evaluation jargon. In the VET sector it includes teaching units, courses, series of courses, curricula, a training or university programme, the services of a vocational training provider, local, regional, national or European VET programmes. Programmes are packages of measures, comprising a succession of activities based on a set of resources, aimed at specific outcomes with defined target groups. A programme comprises a fixed (written) plan or design (programme as plan) and its implementation in practice or conduct (programme as action). Data-based evaluations can describe and assess policies by carving several programmes into evaluands.

pean Commission Action Plan of November 2001, *Making a European area of lifelong learning a reality*, states: 'The Commission, the Member States and the social partners will jointly examine the role and character of voluntary minimum quality standards in education and training' (European Commission, 2002, p. 17). The European Training Foundation writes in its manual *Development of standards in vocational education and training*: 'The creation of market economy structures in these countries often brings with it increased, and frequently completely different requirements in terms of the general abilities, knowledge and skills required by employers at the intermediate qualification level. These requirements are documented in vocational education and training standards' (ETF, 1999, p. 3). In Germany the contribution of the German Institute for International Education Research, *Bildungsstandards als Beitrag zur Qualitätsentwicklung des Schulsystems*, is the subject of widespread debate (Klieme, 2002). The Federal Institute for Vocational Training sees training standards as a central component of a 'new paradigm for the creation of vocational profiles' (Sauter and Schmidt, 2002, p. 21).

Initially, we should specify that the standards presented, which are drafted at the European level and in Germany, refer to VET measures, programmes, training and university courses and portray desirable qualities of the aspects that evaluations describe and rate. From the evaluation perspective we are talking about 'programme standards', or evaluation quality requirements.

This paper deals with the 'evaluation standards' that impose requirements on evaluations themselves. Repeated confusion of these two levels occurs, e.g. when in some evaluation models the programme standards are erroneously labelled 'evaluation criteria', i.e. yardsticks for measuring the merit or the applicability of the programme being assessed.

Since evaluation in VET is a potentially important reference point and has such VET standards (Section 3.3), it is particularly significant for evaluation terminology to clarify the definition of 'standard'. This is vital for clear communication on VET quality between policy-makers, programme managers, social partners and other stakeholders as well as evaluators. Focusing on German and

Anglo-Saxon countries, we provide an excursus on the meaning of the word 'standard' (Section 2.2).

The study addresses the following central questions:

- (a) does Europe need a code in the guise of evaluation standards to ensure and improve the quality of VET evaluations?
- (b) do existing general evaluation standards win the approval of European experts in evaluation and VET?
- (c) what opportunities and what risks are seen in propagating a single set of VET evaluation standards in Europe?
- (d) what cultural and professional values and requirements should such a code address?
- (e) are any standards not applicable to specific VET contexts?
- (f) are there quality requirements for evaluations in VET contexts which are missing in general evaluation standards? Should there be additional standards or extensions of existing standards?
- (g) are the standards equally suited to evaluations in organisations (VET institutions), at the local/regional level (i.e. cooperation of several institutions, schools and enterprises), at the national and European level?
- (h) what recommendations are made in relation to discussing and disseminating standards for evaluation in the evaluation profession, vocational educators and trainers, and in the government VET authorities?

The subject of this paper is the evaluation of European VET programmes and measures. It aims to determine the status of evaluation standards in this area and to present well-founded suggestions for their specification.

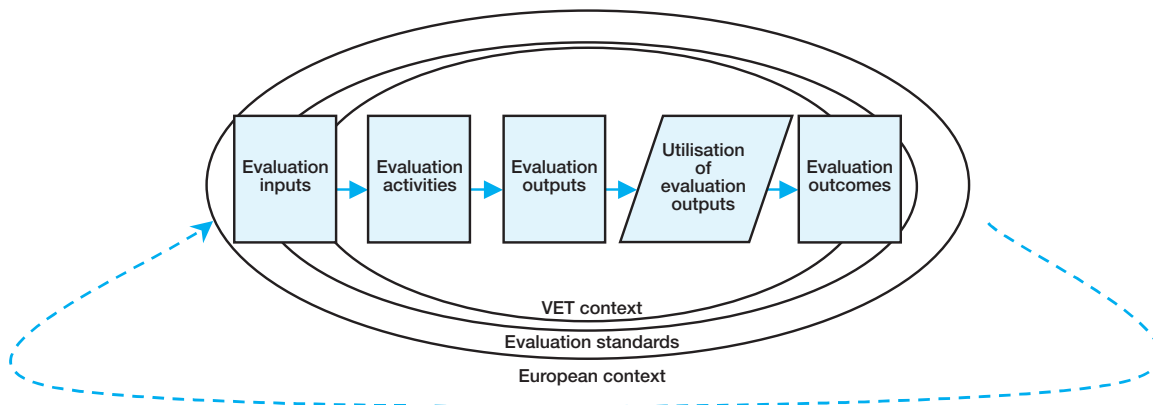
In the following chapters we will survey a wide range of evaluations and diverse elements of evaluations, from their inputs to the output and its utilisation. They will be localised in the VET context and subjected to a critical assessment with the help of evaluation standards, these standards being honed to specific requirements for VET evaluations. The discussion mirrors the background of the authors, and involves experts in evaluation and VET and related European literature.

Chapter 2 presents general – i.e. applicable in all policy fields – sets of evaluation standards as performance processes. Other relevant standards

for certain policy areas and political organisations are presented in Chapter 3. Intercultural transferability and application to VET are also addressed. In the following three chapters, experts speak through three channels: dialogue events on standards (VET experts from Germany and Austria), a

survey of 19 evaluation experts in widely divergent European countries, and critical analysis of recent European literature on VET evaluations. Chapter 7 summarises the findings of the analysis and recommends refinements to standards for VET evaluations.

Figure 1: **Subject of this paper**



Source: Own depiction based on Zorzi et al. (2002)

2. General evaluation standards

This chapter presents the development and context of German, Swiss and US evaluation standards and sketches their goals and composition, using the German code as an example. Subsequently, we explain the basic philosophy of evaluation standards with reference to the content and relationship of the four groups of standards. We then discuss their character as maximum standards which should support dialogue and learning about good evaluation practice. We distinguish between evaluation standards referring to evaluation services and guiding principles that relate to evaluator competence and performance. In conclusion, we trace the connection between evaluation standards and evaluation models.

2.1. Background and purpose of evaluation standards

Professionalisation of evaluation in the US since the mid-1970s has involved the development of various sets of standards to register and control the quality of evaluations. The evaluation standards of the Joint Committee on Standards for Educational Evaluation (JC) are widely known. The JC first published *Standards for evaluation of educational programs, projects and materials* in 1981 (JC, 1981). In 1994 the JC, which by then belonged to the later-founded American Evaluation Society, presented the *Program evaluation standards*. They were revised in a laborious five-year review process. They now go beyond schools and universities. A reference to education and training was consequently only mentioned in the subheading of the publication.

The JC standards were translated into German (JC, 2000) and initially adapted by the Swiss Evaluation Society (SEVAL, 2002). The DeGEval

also decided to base its own standard-setting process on the work of the JC to harness the 20 years of materials and published expertise in related JC standards and to facilitate international exchange. A commission, made up of representatives of various fields of application and academic disciplines, revised the JC standards to match the German and Austrian situation and had them reviewed by qualified commentators. In autumn 2001 the DeGEval (2002) approved the evaluation standards.

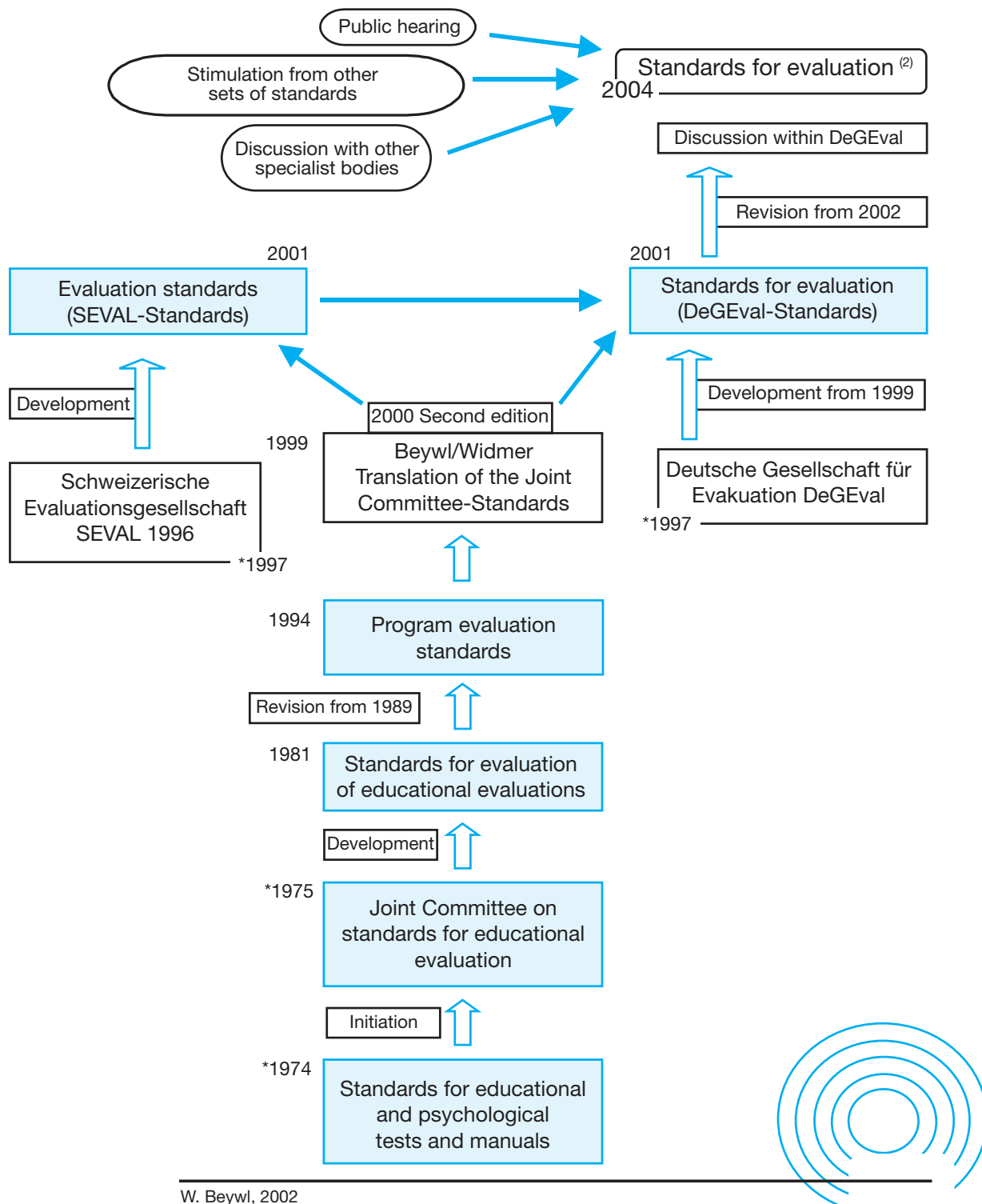
This paper focuses on the DeGEval standards. Their basic philosophy, their systematic organisation, their designation of most of the standards and their use of terminology often adhere to the JC and SEVAL standards. Like the latter, they are helping to adapt the US model to European policy and research traditions. If the DeGEval standards are the starting point for discussion and analysis of applicability to VET, this takes place in the name of the 'family' of evaluation standards, to which the JC and SEVAL standards both belong. Whenever our statements essentially apply to all three sets of standards, we will call them simply 'evaluation standards'.

The evaluation standards address evaluators, individuals and organisations who commission evaluations, and stakeholders in the programme undergoing evaluation and other evaluands. The standards are designed primarily as tools of dialogue and well-founded reference points for evaluations. The standards furnish adequate, appropriate aids for all evaluation phases. The weighting of the standards depends on the main objective of an evaluation. We distinguish between phase-related objectives in the course of the evaluation cycle and cross-sectional tasks, which are performed several times or continually in the course of an evaluation ⁽²⁾.

⁽²⁾ For an overview see DeGEval (2002), pp. 38-41.

Figure 2: Evolution of DeGEval standards

Degeval standards: creation and perspectives



W. Beywl, 2002

Table 1: **Main tasks in an evaluation**

Phase-related tasks	A. Decision on performing an evaluation
	B. Definition of evaluative question
	C. Evaluation planning
	D. Information collection
	E. Information processing
	F. Evaluation reporting
Cross-sectional tasks	G. Evaluation budgeting
	H. Evaluation contract
	I. Evaluation management
	J. Evaluation staffing
Source: author's representation	

The standards are also meant to be interfaces for initial and continuing training in evaluation. They can likewise be employed in the evaluation of evaluations (meta-evaluation) and, finally, make evaluation transparent to the general public as the performance of a profession.

The DeGEval standards consist of 25 *Standards für Evaluation*. Like the JC and SEVAL standards, the DeGEval standards prescribe four basic qualities for evaluations: utility, feasibility, propriety and accuracy. The 25 standards are divided into these four categories ⁽³⁾. These standards, limited to three printed pages, are supplemented by materials, explanatory notes, aids and checklists as well as an annex (DeGEval, 2002) ⁽⁴⁾. A transformation table shows which individual standards from the three related sets correspond and enables users of the less established SEVAL and DeGEval standards to consult the copious body of JC materials (JC, 1994, 2000). To identify the standards unambiguously, we will use the abbreviations listed at the end of this paper ⁽⁵⁾.

2.2. Philosophy of evaluation standards

The four attributes – utility, feasibility, propriety, accuracy – reflect the thrust of the standards associated with each of the four groups. It is to be hoped that an evaluation observes all four criteria.

The accuracy standards in Group 4 underscore the incontestable demand that evaluation be based on scientific methods. They require that the scope of the evaluation and its findings be stated precisely (G1/A1 and G2/A2) and that the procedure and sources of information tapped be presented in a manner conducive to comprehension and verification (G3/A3 and G4/A4). Standards G5/A5 to G7/A7 treat validity, reliability, systematic error checking and qualitative and quantitative data analysis, which are crucial requirements of empirical social science research. G8/A8 stresses that conclusions must clearly follow from the empirical data. Finally, G9/A9 demands that evaluations submit to systematic meta-evaluation.

⁽³⁾ The US JC standards are composed of 30 individual standards, which were partially combined in the DeGEval standards, yielding a set of 25. See Annex 1: transformation table.

⁽⁴⁾ The annotated DeGEval standards can be found in English translation in the annex. We therefore forego a detailed description at this point.

⁽⁵⁾ This publication usually cites the DeGEval standards. To avoid misunderstandings, the reference numbers of both the German and the English texts will be given, e.g. N1/U1 for the *Nützlichkeitsstandard* No N1 or the translated Utility Standard No 1. In exceptional cases we will also cite the JC standards. They will be indicated as follows: JC-U1 for Joint Committee Utility Standard No 1.

The third group, propriety standards, contains requirements which we know from the ethics of science (F2/P2): protection of individual rights, F5/P5; disclosure of findings; and additional demands which result from the clash between evaluation as assessment of social practice and as scientifically based procedure (F1/P1, formal arrangements; F3/P3, complete and impartial review; and F5/P5, disclosure of findings).

The second group, feasibility standards, emphasises that in implementing evaluations – in contrast to basic scientific research – one must always consider economic, social, political and organisational factors which impinge on the programmes, etc., to be evaluated. Standards D1/F1 to D3/F3 state that compromises and adaptations must constantly be made. Procedures must be appropriate to the practice which is to be described and evaluated. They must be introduced and performed diplomatically. They must be efficient in terms of their cost-benefit ratio to be accepted by practitioners and to be politically viable.

The first-mentioned group of standards uses 'utility' to label the central goal of evaluations and suggests that the information and conclusions they provide should actually be used by the evaluated programme stakeholders (N8/U8). Analysis of, and research on, evaluation have derived seven requirements which utilisation and worth of evaluations must demonstrate: identified and adequately involved stakeholders, clarified evaluation purposes, credible and competent evaluators, suitable selection of data, transparently presented values, complete and clear reporting, and timeliness of evaluation activities.

It may seem odd that the utility standards come first and the accuracy standards last in the set. This is no indication of their relative status. It highlights the often unresolvable conflict between scientific merits and the requirements of evaluation users which frequently arise in the course of evaluations. 'In practice, therefore, the evaluator must struggle to find a workable balance between the emphasis to be placed on procedures that help ensure the validity of the evaluation findings and those that make the findings timely, meaningful, useful to the consumers' (Rossi et al., 1999, p. 31) ⁽⁶⁾. The systematic

listing of the four groups in the DeGEval, SEVAL and JC standards underscores the fact that struggling for an appropriate balance between differing criteria, sometimes diametrically opposed, evaluation quality is at the heart of good evaluation in theory and practice.

The outline of the evaluation standards does not connote any weighting, neither among groups of standards nor between individual items. Widmer (2000) states that the fact that each group contains a different number of standards does not permit us to draw any conclusions about the relative importance of any group. Weighting of individual standards should be conducted for each separate evaluation, taking account of its determinants. This is very significant because individual standards sometimes lay competing claims. It is the job of the evaluator to decide which standards to prioritise, to state this expressly and justify the choice. Evaluators are always involved in a bitter tug of war between two or more sides.

Evaluation standards are designed to unfold and explain the broad spectrum of quality norms and bring them to the attention of those concerned. Different quality criteria should not be played off against each other. They should serve as signposts for careful planning, conducting and analysis of evaluations.

2.2.1. Excursus on the meaning of the word 'standard': minimum vs. maximum standards

The term standard is used in many ways. We will focus on the difference between minimum and maximum standards.

The German word *Standard* is derived from its English cognate: 'yardstick, norm, rule' (<19th century), borrowed from the modern English word 'standard', originally 'flag'. The shift of meaning in English from 'flag' to 'norm' has not been reliably mapped (either via 'guiding', 'gauging' or 'king's standard [royal flag], or as a landmark providing orientation') (Kluge, 1999, p. 787).

Webster (1989, p. 1385) gives a total of 28 meanings for standard(s), including the original meaning, No 12 in the list: 'a flag indicating the presence of a sovereign or public official' and No 13, 'a flag, emblematic figure, or other object

⁽⁶⁾ Whether evaluation is an academic discipline or a scientific profession is the subject of great controversy. The topic was the focus of a workshop at the conference of the European Evaluation Society (EES, 2002).

raised on a pole to indicate the rallying point of an army, fleet, etc.’ The following, non-academic meanings illustrate the versatility of the term:

- (a) an object considered by an authority or by general consent as a basis of comparison; an approved model;
- (b) anything, as a rule or principle, that is used as a basis for judgement;
- (c) an average or a normal requirement, quality, quantity, level, grade, etc.;
- (d) standards: those morals, ethics, habits, etc., established by authority, custom or an individual as acceptable;
- (e) the authorised example of a unit of weight or measure.

Webster’s differentiation between standards and criteria provides input for discussion of the use of the two terms in the language of evaluation: ‘A “standard” is an authoritative principle or rule that usually implies a model or pattern for guidance, by comparison with which the quantity, excellence, correctness, etc., of other things may be determined. [...] A “criterion” is a rule or principle used to judge the value, suitability, probability, etc., of something, without necessarily implying any comparison.’

The last-quoted definition of standard shows that it can be used for comparison with some specified quantities as well as with less operational items such as ‘excellence’ ⁽⁷⁾. For the highest possible terminological clarity, we use the universal distinction between the two extreme varieties of standards (which should apply to both evaluations and evaluands). A ‘minimum standard’ states, usually in rather technical terms, specific (ideally quantitatively operationalised) minimum requirements, which must be strictly observed (here: by an evaluation) so that high quality can be ascribed to it. A ‘maximum stan-

dard’ states, usually in lay terms, which leave scope for interpretations, the envisioned ideal (which an evaluation should fulfil to be judged to be of high quality).

In consultation with evaluation experts, who complemented our study, we noted that the term ‘standard’ possesses very different connotations, depending on national origin, academic background and one’s role in the evaluation (commissioner/evaluation team):

- (a) colleagues from the United Kingdom primarily associate standard with quantified, unconditionally binding minimum standards (as in a British Standards Institute definition, for example). ‘A standard is a published specification that establishes a common language, and contains a technical specification or other precise criteria and is designed to be used consistently, as a rule, a guideline, or a definition [...]’ ⁽⁸⁾;
- (b) psychologists – at least those who are statistically inclined – usually think in terms of minimum standards, while sociologists tend toward maximum standards;
- (c) commissioners (particularly if they have introduced quality management systems) ⁽⁹⁾ often prefer operationalised minimum standards, e.g. stipulated in requirement specifications, while evaluators favour maximum standards because they guarantee the necessary flexibility for planning and conducting evaluations.

Because of such ambiguities, the drafters of the SEVAL and DeGEval standards considered replacing the term standard with another such as ‘norm’, ‘code’ or *Richtlinie*. These evaluation societies decided differently, however, because such terms are also ambiguous from discipline to discipline and would not increase clarity. Our experience shows that a consensus can only be

⁽⁷⁾ Harvey and Green (1993) for the five quality dimensions of human services and their fundamentally varying capacity for being put into operation.

⁽⁸⁾ This layperson’s definition is given on the BSI education sites at <http://www.bsi-global.com/Education/index.xalter>

⁽⁹⁾ The Standards Policy Team of the Regulatory Affairs and Standards Policy Directorate, Industry Canada stresses maximum standards in its first definition section, whereas in a second section it defines minimum standards with reference to the International Organisation for Standardisation (ISO). ‘A standard is broadly defined as a publication that establishes accepted practices, technical requirements and terminologies for diverse fields of human endeavour. The International Organisation for Standardisation (ISO) defines standards as documented agreements containing technical specifications or other precise criteria to be used consistently as rules, guidelines, or definitions of characteristics, to ensure that materials, products, processes and services are fit for their purpose.’ Available from Internet: <http://strategis.ic.gc.ca/SSG/sp00447e.html#NSS> [Cited 13.11.2003]. The German counterpart of the BSI is the *Deutsches Institut für Normung*, the Austrian is the *Österreichisches Normungsinstitut* and the German areas of Switzerland have the *Schweizerische Normenvereinigung*. *Norm* is closer in meaning to the international and English word ‘standard’ than to the German word *Standard*.

achieved through widespread intensive perusal or trial application of the evaluation standards. Faithful to the JC tradition, SEVAL and DeGEval chose to retain the term standard.

In this paper, we have chosen to state expressly each time whether we mean maximum standards (as in the JC, SEVAL and DeGEval standards) or minimum standards (as in quality management).

The evaluation standards discussed in this paper are conceived as maximum standards. An ideal evaluation would adhere to each individual standard that is theoretically applicable to this evaluation. The JC standards expressly provide for the possibility of *a priori* non-applicability of certain standards to a concrete evaluation project ⁽¹⁰⁾. This ideal, already qualified, can rarely be achieved in practice unless the requirements of two or more standards prove to be contradictory or financial resources do not suffice to meet all standards ⁽¹¹⁾. Even though a specific evaluation can hardly comply with all standards equally, evaluators should strive to take account of each – where applicable – as far as possible.

European VET discussion involves various standards. In the introduction we termed them ‘programme standards’ in contrast to the ‘evaluation standards’ covered in this paper. Typically VET standards are copious bodies of rules. For example the German term *Ausbildungsordnung* (training regulation) has more recently been translated as ‘VET standard’. The government-issued training regulations in Germany dictate requirements for ‘state-recognised training occupations that require formal vocational training’ (Sauter and Schmidt, 2002, p. 7). The 1999 publications of the European Training Foundation aim to create a similarly

comprehensive body of standards to support eastern European countries in developing VET standards. Another prominent example of a detailed, descriptive standard containing definitions, specifications, checklists, codes, the reasoning behind them and much more, is ISO 9000:2000 comprising approximately 40 printed pages.

However, in this paper we use standard as a label for short, succinct texts, often limited to one sentence, and rarely exceeding three ⁽¹²⁾. These maximum standards are statements for evaluation planning and execution. They constitute a basis for meta-evaluations.

2.3. Standards for evaluations and guiding principles for evaluators

In the US we find, apart from the JC standards, the *Guiding principles for evaluators* (Shadish et al., 1995) ⁽¹³⁾. The latter were developed as professional guidelines or codes of ethics by the American Evaluation Association ⁽¹⁴⁾.

While the JC, SEVAL and DeGEval standards refer to the quality of evaluations as a service, the American Evaluation Association guiding principles state requirements of professionals who plan and conduct evaluations, i.e. evaluators, occasionally also of commissioners ⁽¹⁵⁾. While the former focus on the quality of rendering the service, the latter concentrate on evaluators’ professional and personal skills and their adherence to general laws and codes of ethics and assumption of professional and personal responsibility ⁽¹⁶⁾.

⁽¹⁰⁾ For example, in the *Checklist for applying the standards* (JC, 1994, p. 18 f). It is also clearly stated in the analogous checklist attached to the DeGEval standards.

⁽¹¹⁾ Examples of non-applicability and non-achievability of applied standards are given in Section 5.1 of the currently unpublished meta-evaluation by Jenewein (2001). For an example in VET, see Section 5.1.

⁽¹²⁾ The DeGEval standards make a clear distinction between standards and explanatory notes. JC and SEVAL publications do the same thing. The standard *per se* is the ‘presentation of the standard in the form of a should statement’ (JC, 1994, p. 7). ‘(The standards [...] comprise a term and a description in one sentence’ (SEVAL, 1999, p. 2).

⁽¹³⁾ Their relevance for the American Evaluation Association is evidenced by the fact that these guiding principles are printed verbatim on the initial pages of each issue of the *American Journal of Evaluation*.

⁽¹⁴⁾ The American Evaluation Association and 14 other organisations were involved in elaborating the JC standards.

⁽¹⁵⁾ The terms ‘standard’ and ‘guiding principle’ are not mutually exclusive and, ultimately, they are chosen arbitrarily. We propose the convention of using ‘standards’ for evaluation services and ‘guiding principles’ for evaluators, cf. Section 3.1.

⁽¹⁶⁾ Further examples of the category ‘guiding principles’ are the *Guidelines for the ethical conduct of evaluations* of the Australasian Evaluation Society (1998), which address commissioners, users and teachers in the field of evaluation, and the *CES Guidelines for Ethical Conduct* of the Canadian Evaluation Society. A more comprehensive discussion and a comparison are found in Beywl and Widmer (2000).

The guiding principles are much more general and broader than standards. Sanders (1995) finds no contradictions or inconsistencies between the guiding principles and the JC standards (Sanders, 1995; pp. 50-51). The former concentrate on evaluators' professional values, whereas the latter focus on professional performance.

There are five guiding principles. Systematic enquiry is basically contained in the accuracy standards and in JC-U3, Information scope and selection. The guiding principle Competence matches JC-U2, Evaluator credibility, and JC-A12, Meta-evaluation. Integrity and honesty principles are found in the feasibility, propriety and, to some extent, accuracy standards.

The fifth guiding principle, Responsibilities for general and public welfare, touches on several standards from the four groups of JC, SEVAL and DeGEval standards. JC-U1, Stakeholder identification, specifies 'the general public' as a potential stakeholder requiring consideration. JC-D2/F2 addresses political viability. 'Evaluations are politically viable to the extent that their purposes can be achieved with fair and equitable acknowledgment of the pressures and actions applied by various interest groups with a stake in the evaluation' (JC, 1994; p. 71). In the Propriety group, JC-P6 elevates disclosure of findings to a central quality criterion. Finally, Accuracy standard JC-A12 requires meta-evaluations '[...] (which) should enhance the credibility of particular programmes evaluations and the overall evaluation profession' (JC, 1994, p. 185).

The public welfare duty ('evaluators have obligations that encompass the public interest and good') establishes a peculiarity of the guiding principles which was debated most ferociously during their drafting ⁽¹⁷⁾. The same goes for the promulgation of the principle 'freedom of information is essential in a democracy.' This can be viewed as a vote for making publication of evaluation reports obligatory.

The JC standards, and even more the DeGEval standards, are more reluctant than the guiding

principles to express requirements based on such codes of ethics or the theory of democracy. One reason is that evaluation standards are typically drafted by a team including evaluators and commissioners, and the conflicts of interests between these groups, e.g. on obligatory publication, already lead to compromise solutions at the early stage of negotiations. In contrast, evaluator professional organisations are much freer in formulating guiding principles. They can stipulate further voluntary obligations.

Since evaluations of VET programmes and measures in the EU and its Member States are set in an intricate stakeholder mesh ⁽¹⁸⁾, existing evaluation standards are suitable starting points for a discussion of VET evaluation standards. This debate may stimulate evaluators active in trade and professional associations in Europe to refine guiding principles.

2.4. Evaluation standards and models

Evaluation standards are designed to be suitable for a huge variety of evaluation approaches and to be applicable to the broadest possible scope of applications. They are generally appropriate for both formative evaluations, which accompany the shaping of the evaluand and attempt to foster improvements, and for summative evaluations, which calculate a balance, usually on one evaluand.

In past decades evaluation models of all shapes and sizes have emerged ⁽¹⁹⁾. They differ, in particular, in their epistemological foundations, the academic field of their authors, the incorporation of social values and interests, participation and use conceptions, evaluation purposes, advance organisers, relationship of the evaluation to the phases of the programme concerned, stressed dimensions of the evaluand and methodological preferences.

Sometimes we encounter developers or users of a certain model who assume that this is the

⁽¹⁷⁾ See various articles in the special issue of *New directions in program evaluation*, No 66. San Francisco: Jossey Bass, summer 1995.

⁽¹⁸⁾ Cited examples are the legislative and central government executive branches, employers, unions, and professional teacher and trainer associations.

⁽¹⁹⁾ A survey is found in Beywl (1988), Owen and Rogers (1999), Russon, C. and Russon, K. (2000), Stufflebeam (2001), Kellaghan and Stufflebeam (2002).

best or even the only applicable evaluation model. They then equate their brainchild with evaluation. This may result from a narrow, subject-oriented perspective or from institutional embedding of evaluation tasks in a national or international organisation or agency. It may also be related to the intention of jockeying one's own model into a more favourable bargaining position in negotiating evaluation policy ⁽²⁰⁾.

JC, SEVAL and DeGEval standards claim to cover the entire spectrum of evaluation models and incorporate pluralistic epistemology and methodology. On the one hand, they do not favour any specific evaluation model or group of models. On the other hand, it has been shown that some models, especially if they are used 'purely', conflict with some standards ⁽²¹⁾. In practice a mix of evaluation models is applied when drawing on evaluation and analysis experience to design and implement a concrete evaluation. In so doing, evaluators often meet evaluation standards, even if they do not know them. This is not to say that all, or even the majority of, evaluations are high quality in terms of evaluation standards; judging this requires systematic meta-evaluations, which have not yet been conducted ⁽²²⁾.

In this report we cannot provide a systematic survey of evaluation models. Patton (1997, p. 192) lists 57 approaches in a table. Each year anthologies or textbooks introduce a new variant or an entirely new approach ⁽²³⁾. Evaluation

models in English dominate. Most of them are from the US. A few Continental approaches have also found a foothold or promise to add a new dimension to evaluation theory and practice (e.g. Pawson and Tilley, 1997, Kushner, 2000) ⁽²⁴⁾.

The following survey outlines a few of the most prominent evaluation models employed in widely divergent fields, including VET. The depiction is organised in terms of value interpretation, which standard N5/U5 stresses. This corresponds to the notion that evaluation takes values ⁽²⁵⁾ as a constituent reference point of practice.

The following outline of the four main types is succinct and is no substitute for thorough analysis ⁽²⁶⁾. Categorisation is guided by the evaluation model's consciousness of values ⁽²⁷⁾. Commonly we find overlaps between categories, which result from ambiguities in model descriptions, particularly when the subject of values is only treated implicitly.

Value-distanced approaches follow the tradition of thinkers such as Max Weber or Karl Popper and eliminate value judgements from the evaluation process. Theoretical framing of an evaluation and implementation in empirical investigations operate 'objectively' according to strict rules; the utilisation of evaluation findings is delegated to the external public democratic process ⁽²⁸⁾.

Value-positioned approaches expressly assume that societies are marked by stark power imbalances and social and economic inequality. Evalu-

⁽²⁰⁾ In the 1990s, Germany often saw a monopoly claim to total quality management for large areas of human services. The relation between quality management and evaluation, which are equally represented in VET contexts, has not yet been studied in international comparison.

⁽²¹⁾ Stufflebeam (2001), who performs a systematic comparison of a total of 22 evaluation models across the entire board of JC standards, speaks in these cases – rather argumentatively – of pseudo-evaluations and quasi-evaluations.

⁽²²⁾ But this is not the case for Widmer's (1996) meta-evaluations in a wide range of Swiss policy areas.

⁽²³⁾ For example, Mark et al. (2000), Kushner (2000), Hale (2002).

⁽²⁴⁾ No assessment which systematically and comparatively presents the evaluation approaches developed in Europe outside of Ireland and UK has been published.

⁽²⁵⁾ It is beyond the scope of this paper to define the multifaceted term 'value'. An intercultural comparison reveals that North American evaluation literature often uses 'value' in collocation with 'material', 'social', etc. (Cf. explanation of JC Standard U4).

⁽²⁶⁾ An initial systematic portrayal is found in Beywl et al. (2003), dealing primarily with evaluations of poverty avoidance and social inclusion policies and programmes. A comparative study and survey of evaluation models focusing on VET is not yet available. A first approximation can be consulted in the annotated bibliography by Beywl and Schobert (1999).

⁽²⁷⁾ Assignments are not performed analytically, by maintaining, for instance, that cost-benefit analyses are bound *ipso facto* to the value judgements of shareholders (a stakeholder subgroup) or that goal-free evaluations mainly reflect values that are widespread in society (thus confirming the value hierarchy). Such mutually critical analyses form the nucleus of the 'paradigmatic debates' in evaluation methodology (Guba and Lincoln, 1997; Pawson and Tilley, 1997; *Philosophies and types of evaluation research* authored by Eliot Stern in this publication).

⁽²⁸⁾ VET evaluations tend to take place in enterprises managed as meritocracies. This would require a fundamental adaptation of evaluation models associated with 'open' and 'experimental society'. We believe this is a current research objective.

Table 2: Exemplary models of evaluation by value interpretation

Model family	Model Type	Models	Author ^(a)
Value-distanced	Effectiveness-oriented evaluation	Goal-oriented effectiveness estimation	Madaus and Stufflebeam (1988)
		Experimental impact model	Shadish et al. (2002)
		Quasi-experimental impact model	Heckman and Smith (1996)
	Efficiency-oriented evaluation	Cost-benefit analyses	Levin and McEwan (2001)
	Result-oriented evaluation	Goal-free result assessment	Vedung (1999)
	Programme-theory-oriented evaluations	Theory-driven evaluations	Chen (1990)
Value-positioned	Participative evaluation	Empowerment evaluation	Fetterman (2000)
		Democratically balanced evaluation	House and Howe (2000)
Value-prioritising	Stake-oriented evaluation	Decision-oriented evaluation	Stufflebeam et al. (1971)
		Utilisation-focused evaluation	Patton (1997)
Value-relativistic	Constructivistic evaluation	Responsive evaluation	Guba and Lincoln (1989)
^(a) Here we cite either creators of the evaluation strategy or authors who give a well-founded overview of the given evaluation model			

ations should counterbalance the value hegemony in the political and cultural spheres by strengthening the weak and giving them an audible voice in the political process.

Value-prioritising models also assume strong disequilibria in society, but thus restrict themselves to making them transparent and accessible to the negotiation of particularly relevant/socially accepted values. For instance, they may demand involvement of all stakeholders in the determination of questions and discussion of findings and may work toward prioritisation and a minimum consensus.

Value-relativistic models underscore the dominant significance of values in planning, executing and utilising evaluations. They detect value conflicts in all phases and maintain existing tensions without taking sides. Motivation and social energy in using evaluation findings derive

from consciously and publicly stated differences in values and interests among stakeholders.

The explicit reference to evaluation models in conception, and particularly in written reporting, of evaluations offers an opportunity to assess the suitability of certain approaches for concrete VET evaluation tasks, to criticise them and contribute to refining evaluation methodology. An evaluation standard could demand specification and justification of the model (or the two or more models) which were used to design an evaluation and to explain why it/they fit the given evaluation purpose, the evaluation questions and the specific VET external variables. Such a disclosure and justification requirement would encourage propagation of evaluation models in Europe, discussion of their weaknesses and strengths and development of an awareness of the need for meta-evaluation (recommendation 7).

3. Transferability of standards

This chapter discusses the general question of how transferable evaluation standards originating in the US are to the European social, political and cultural context. It goes on to present the measures undertaken by the EU and its Member States to develop independent evaluation norms, including some international sets of standards. The chapter then conveys some initial perceptions on the standards' transferability to vocational training. Later chapters will expand these ideas.

3.1. Intercultural transferability of standards

As demonstrated in the previous chapter, the development of evaluation standards in Europe was stimulated by the US JC standards ⁽²⁹⁾. At first glance this seems a good idea because of the high costs of devising standards, but it must also be seen in the light of a general dominance of the US evaluation approach. As already mentioned, Europe has developed hardly any independent evaluation models of its own. This suggests that evaluations on this side of the Atlantic largely follow the American lead. Vedung (1999) is an exception for an evaluation approach developed in Europe ⁽³⁰⁾. Pawson and Tilley (1997) describe an evaluation model which explicitly espouses European traditional thinking and represents a deliberate departure from the US precedent ⁽³¹⁾.

Professional standards are usually shaped by values and norms, which can vary widely from culture to culture. In addition, the configuration of the parliamentary system is an important determinant of national evaluation culture ⁽³²⁾.

In his 1986 publication, Stufflebeam, the long-serving chairman of the Joint Committee, claims that the JC standards have limited use outside the US. He writes that other countries have adopted adaptations of the standards. Few would question the transferability of standards based on procedures derived from social sciences, the 'accuracy' category ⁽³³⁾. These cross-cultural norms have been formulated almost identically in very different fields in the US and in European countries, for example in the *British psychological society's code of conduct* and in the British Sociological Association's *Statement of ethical practice*. In addition, quality standards exist for certain parts of programme evaluation. These include standards for the design of experimental research and objectivity, reliability and validity specifications as survey quality criteria. For information on the formal aspects of quality assurance, we refer to quality management concepts such as the ISO/EN/DIN 9000ff norms ⁽³⁴⁾.

In North America, where there is a profound mistrust of State control in general, independent assessment seems a more logical approach. The public expects to be informed of the costs and benefits of government activities and the US has long dedicated considerable resources to

⁽²⁹⁾ They have been implemented in countries with extremely different evaluation cultures, such as Brazil and Israel. A European example is Sweden (Marklund, 1984).

⁽³⁰⁾ Vedung (1999) is available in Swedish, English, German and Spanish. See also Beywl and Taut (2000).

⁽³¹⁾ Cf. the detailed description in this report (Eliot Stern).

⁽³²⁾ 'Competitive democracies' (dominance of the majority principle) and 'consociational democracies' (consideration of all relevant interests sometimes going as far as the principle of unanimity) (Jesse, 1993) tend to assign different functions to the evaluation of political programmes and measures. The rapid development of an evaluation culture in Switzerland, the home of consociational democracy, may indicate that independently obtained evaluation findings foster amicable resolution of conflict and willingness to compromise. It would be interesting to analyse national VET evaluation cultures as a function of the respective political system, incorporating structural characteristics of the VET systems.

⁽³³⁾ See also the discussion on accuracy standards in Section 6.4.

⁽³⁴⁾ Beywl and Schobert (2000) give an overview of the relationship between evaluation and quality management in vocational training. Speer (2001) compares and differentiates between evaluation and benchmarking in company personnel management.

academic evaluations of training and labour-market programmes. In Europe, training as part of labour-market interventions is a much more recent tool, especially in southern EU countries. Since programmes in this field are a relatively new phenomenon, there is a dearth of econometric data and programme designs. The US has a far larger reservoir. Schmidt (2000, p. 427) points out the striking absence or infancy of social science experimentation in Europe. The differing evaluation cultures in North America and Europe must be taken into account.

A survey of 1645 companies in Finland, Germany, Ireland, Northern Ireland and the UK identified differences in the evaluation of training activities (Field, 1998a). The UK conducted more training evaluations than the other countries; this was particularly evident for evaluations of pre-training activities and reflective evaluations. In Germany comparatively little evaluation takes place during training. Finland conducts a relatively large number of evaluations immediately after training courses finish. The countries in which reviews are carried out most often evaluate training as soon as courses end. Next most frequent are evaluations before training starts, followed by evaluations after participants have returned to their jobs. Evaluations are least common during training activities. The purposes of the evaluations vary in focus correspondingly. One often-mentioned aim is to test whether training has fulfilled its objective. Evaluations which concentrate on improving participants' abilities to perform their job are identified as important. This was especially true in the UK. Thus differences of emphasis characterise the evaluation practices of various European countries. However, this does not affect evaluation standard enforcement options ⁽³⁵⁾.

Those standards which demand a high level of social awareness during planning and management of evaluations in the national/regional context are likely to be sensitive in intercultural applications (Rost, 2000). This applies especially to the following standards:

- (a) identification of stakeholders in the evaluated programme (N1/U1);
- (b) relative significance of personal, social and evaluand-related skills for the credibility of evaluators (N3/U3);
- (c) disclosure and discussion of values and interests as the basis for judgements (N5/U5);
- (d) anticipation of the various positions to ensure their advocates' cooperation and to prevent deliberate obstruction of the project (D2/F2);
- (e) consideration of the culturally determined and legally protected inalienable personal interests of all those involved in the evaluation (D2/F2/P2);
- (f) attempts to ensure all relevant interests are treated fairly (F4/P4);
- (g) publication of findings (F5/P5).

It would be useful to test these assumptions through empirical research, but this may result in difficulties, since professionalisation of evaluation and the emergence of evaluation cultures are recent developments in Europe. Often the standard sets are not sufficiently known, particularly their details. This makes it difficult to conduct surveys to gather statements and critical comments on individual standards and their cross-cultural applicability. To offset this barrier, the empirical part of the investigation uses a mix of group discussions (combined with presentation of evaluation standards), an electronic survey of experts (which assumes a certain degree of familiarity with evaluation standards), and content analysis of current European literature on the subject (a non-interactive process). However, the components of this pilot study are no substitute for an analysis of intercultural transferability. Future studies will have to resolve this issue ⁽³⁶⁾.

Since some standards refer to several aspects of desirable evaluation quality, it is also conceivable to weight the focuses of these standards differently in various European countries. Some standards may be crucial, while others may be meaningless because they are rarely fulfilled in the given cultural context or are nearly always met anyway. Standard N7/U7 'evaluation timeli-

⁽³⁵⁾ The concept of controlling in corporate continuing training, which is partly related to evaluation, was also applied in almost exactly the same way in Germany, the Netherlands and Austria (BIBB et al., 2001).

⁽³⁶⁾ We highly recommend workshops and meetings which benefit from a systematic data collection procedure as an appropriate test of intercultural compatibility. They may contribute to the further development of European-level evaluation standards (Recommendation 10).

Table 3: **Particularly culturally sensitive DeGEval standards**

No	Standard English term
U1	Stakeholder identification
U3	Evaluator credibility and competence
U5	Transparency of values
F2	Diplomatic conduct
P2	Protection of individual rights
P4	Unbiased conduct and reporting
P5	Disclosure of findings
Source: author's representation	

ness' can serve as an example. Its relevance can be judged entirely differently from culture to culture. One society may view strictly designated deadlines as evidence of the contractor's low social status whereas another may make the ability to fulfil deadlines an automatic prerequisite for winning an evaluation tender. Other standards may prescribe behaviour that is entirely natural in certain cultural contexts and yet completely alien to others. They would, therefore, be superfluous or incomprehensible respectively (F5/P5: disclosure of findings). This can lead to serious conflict within multinational evaluation teams or during evaluations of programmes implemented in several countries.

The European Commission observed that in the years 1997 to 2000 an evaluation culture emerged with the following characteristic (Schmitt von Sydow, 2001, p. 9) ⁽³⁷⁾: the majority of the evaluations are mid-term, the rest are *ex post* evaluations. *Ex ante* evaluations are rare. Stakeholder orientation is not a priority of European Commission evaluations. They are formative rather than summative. The white paper points out that it is still too early to speak of a general European Commission evaluation culture. The white paper names several purposes of evaluation (idem, pp. 20-21): '[...] to enhance democratic accountability, to assist political decisions about legislation, policies and programmes, to promote closer understanding between stakeholders and to support the implementation and management of existing programmes'. General

rules or standards are regarded as more effective for maintaining objectivity and neutrality than, for example, any new, formally independent evaluation functions within the European Commission. Standards for the evaluation process can increase the credibility of evaluators (idem, p. 38, Annexe IV). Rules like this would tighten methodology and data reliability (idem, p. 35, Annexe IV). It was also decided that evaluations should not be restricted to the perspective of single directorates but should pose and answer cross-sectoral evaluation questions. Evaluations should be designed as inputs for annual decisions on policy priorities.

The evaluation culture of southern EU Member States is strongly shaped by their obligation to justify structural appropriations (European Commission, 1999b, Vol. 1, p. 45); Greece, Spain and Portugal rarely conduct evaluations unrelated to structural funds. In contrast, Denmark, Germany, France, the Netherlands, Sweden and the UK carry out many evaluation activities not related to structural funds. It is not surprising that some of the latter States see evaluations as a part of their political culture and as an expression of the democratic process while the southern European countries often regard evaluations as a chore imposed on them from outside. However, the evaluation activities conducted in the context of EU programmes have accelerated the creation of additional evaluation resources in countries like Germany and France (European Commission, 1999b, Vol. 1, p. 46). A third group of countries including Belgium, Ireland, Italy (Northern), Luxembourg, Austria and Finland, (i.e. mostly smaller, developed countries) predominantly regard evaluation as improved management of public intervention.

We can identify differences between these various evaluation practices which probably result from varying cultural and institutional traditions. Northern Europe is ascribed a parliamentary-democratic evaluation culture (European Commission, 1999b, Vol. 1, p. 202). Wollmann (2002, p. 5 f.) and Vedung (1999, p. 70 f.) claim that Sweden, considered the European leader in evaluation research, has a consensus-oriented political style moulded by parliamentary commissions. These commissions often award contracts

⁽³⁷⁾ This white paper involved 27 European Commission employees from different directorates (members of Working Group 2b) and 18 external evaluation experts from various European countries participating in four hearings.

for studies or evaluations with political relevance. Wollmann writes that the contract recipients are usually university social scientists. In contrast, the EU primarily commissions external bodies to conduct evaluations and carries out very few internally. Wollmann adds that private consultancy firms have the lion's share of the market for external evaluations. He distinguishes between central-level evaluations, whose evaluands are whole programmes, and evaluations of national programmes, which are usually conducted by national (private) evaluation institutes, except in Spain where they are undertaken by universities. On the basis of a study he conducted, Leeuw (2000) concludes that the market for evaluations is a growth industry. The demand for evaluations seems to be expanding more quickly at EU level than at national or regional level.

Because of various institutional arrangements and differently developed evaluation markets, some standards may be particularly culturally sensitive. In another context, Smith et al. (1993, p. 12) identified a fundamental cultural difference in the use of evaluation standards. 'The concept of standards as employed in the US is much less relevant within the Maltese and Indian traditional cultures. Although standards may be imposed from the outside, indigenous standards are unlikely to emerge.'

The discussion of individual standards in Chapter 6 provides details of certain areas of intercultural sensitivities.

3.2. Current approaches to development of evaluation standards in Europe

The institutionalisation of evaluation in the form of evaluation societies is a very recent development in Europe. Societies exist in Denmark, France, Germany, Italy, Spain, Sweden, Switzerland, the UK and Wallonia. Europe also has its own evaluation body, the European Evaluation Society ⁽³⁸⁾. In some European States there has been a critical

look at US evaluation standards. To test the transferability of US standards to Europe, the authors looked at the acceptance of American standards in European Countries. Some national societies have designed or adopted their own standards. The authors contacted them as part of this study if we could not find sufficient information about their standards on their websites, and asked them about the current status of their discussion on standards.

German/Austrian and Swiss standards follow the example of the US standards. The *Société Française de l'Évaluation* is currently developing its own independent standards (SFE, 2002). It has not yet fixed these standards, but internal discussion has reached an advanced stage. In contrast to the JC standards, the French discussion is focusing on the social usefulness and public interest (*utilité sociale et intérêt général*) of the evaluations. It also values the principle of honesty (*principe d'honnêteté*) ⁽³⁹⁾. Referring to product quality policy the *Société Française de l'Évaluation* draft includes guidelines for the structure of evaluation reports and rules on their readability. The French draft is very precise on this point ⁽⁴⁰⁾. The commissioners are responsible for external process management of evaluations and should be directly involved. For example, those responsible for the evaluation should support the development of an evaluation culture in the organisation concerned (IV-6 *Culture d'évaluation*). The *Société Française de l'Évaluation* continues to debate how much attention should be paid to French idiosyncrasies.

The Italian *linea guida per un codice deontologico del valutatore* focuses on the evaluators ⁽⁴¹⁾. It clearly stresses their overriding responsibilities. The contents of the majority of the DeGEval, JC and SEVAL propriety standards feature in the *linea guida*. It is little known in Italy, probably because of the relatively small evaluation market. The Italian Evaluation Society is also considering augmenting the *linea guida* with its own standards (Bezzi, 2002).

The Finnish Evaluation Society also recently developed its own standards (FES, 2002). They clearly focus on 'truth' and 'community'. Such

⁽³⁸⁾ For an overview see Toulemonde, 2000; p. 355. The DeGEval website features a constantly updated link list. Available from Internet: <http://www.degeval.de/weltweit.htm> [Cited 29.10.2003].

⁽³⁹⁾ The 'propriety' standards in the US version do correspond to the term *honnêteté*, but the definition of 'propriety' is much more objective than the ethical appeal the French standards make.

⁽⁴⁰⁾ The US original is also very detailed, in contrast to the German and Swiss versions.

⁽⁴¹⁾ This corresponds to the *Guiding principles* of the US Evaluation Society (Section 2.3).

standards resemble ethical precepts. This seems to be an important consideration for the Finnish evaluation community and its mentality and reflects the origins of the Finnish standards. State institutions played a major role in their establishment. This is not the case for the other national evaluation standards and has obviously influenced the Finns' alternative approach.

The United Kingdom Evaluation Society's *Guidance for good practice in evaluation* (UKES, 2002) focuses on the evaluation process, particularly on cooperation and consultation between the various interest groups. It contains an individual section for each main stakeholder group involved in evaluations: evaluators, commissioners, participants. It also provides guidance and information for participants in self-evaluations. This distinction is not made in any other set of standards. Furthermore, the UK standards contain phrases such as 'it would be helpful', less binding than the prescriptive *sollen* (should) of the DeGEval standards. The guidelines are still the subject of internal negotiations and have not yet been finalised.

Standards for Europe exist alongside those of various national evaluation societies. They resemble the DeGEval standards but are designed for other policy areas than VET, such as development aid. One example is the Danida standards (Danida, 2001). Codes of different national professional organisations are also available and can overlap with evaluation. They are not discussed here but Beywl and Widmer (2000) provide a comprehensive survey.

The European Commission has its own guide and the International Labour Office (ILO) has guidelines which may be relevant for VET in Europe. The following paragraphs describe these publications.

Evaluating EU expenditure programmes: a guide was financed by Directorate General XIX. It was conceived as an aid for evaluating many different kinds of evaluands, including VET programmes and projects with entirely different contexts and contents, and so can be considered pertinent. It identifies the key issues of evaluations as relevance, efficiency, effectiveness, utility and sustainability (European Commission, 1997, p. 18). One of the guide's main focuses is evaluation management and preparation. Selection of evaluators is one part of this. The guide is very detailed and comments on many aspects of eval-

uation which also feature in the standards, but in the more substantial form of a handbook

Guidelines for systems of monitoring and evaluation of ESF assistance in the period 2000-2006' (European Commission, 1999a) was published by the Directorate General for Employment, Industrial Relations and Social Affairs. European Social Fund programmes often include continuing training schemes. Some of these are the 'training' part of the 'measure of assistance to persons' programme category, and the 'teacher training' and 'creation of training/education curricula' parts of the 'measures of assistance to structures and systems' category. Therefore these guidelines can be classified as directly relevant to VET.

The guidelines stipulate that evaluations should follow the logical framework of intervention. That means that indicators should be used to measure the input, output, outcome and impact of a programme. The guidelines clearly state which indicators should be adopted for each stage of the logical framework. They specify which (quantitative) parameters should be selected and how much data needs to be collected (N4/U4). The guidelines also advocate including collection of qualitative data as part of the evaluation process. The analysis of the evaluation context (G2/A2) should cover the 'operational context' and the 'conditions of implementation'. The guidelines explicitly define certain standards: evaluation timeliness (N7/U7); formal agreement (F1/P1); unbiased implementation (F4/P4); efficiency (D3/F3); and disclosure (D5/P5). Thus, most of the DeGEval standards are included in the guidelines, and some are treated more thoroughly. Evaluation utilisation (N8/U8) reflects the use of findings from the *ex post* evaluation. This particularly applies to indicator definition and evaluation scheduling. Mid-term evaluations should be formative and *ex post* evaluations summative.

The MEANS handbooks (European Commission, 1999b) deal with the entire range of potential evaluands from EU politics. Training and employment are most relevant for VET. So the MEANS criteria, which actually originated in regional politics, have also been implemented in other European Commission General Directorates such as DG Employment. The MEANS handbooks stipulate eight quality criteria ⁽⁴²⁾ for evaluations (idem, Vol. 1,

⁽⁴²⁾ The term 'criterion' is somewhat misleading. The term 'assessment dimension' would be more accurate.

p. 169): meeting needs; relevant scope; defensible design; reliable data; sound analysis; credible results; impartial conclusions; and clear report. These are explained in detail, corresponding largely to the specifications of the DeGEval standards and their US predecessor. The MEANS handbooks focus on the 'workmanship' of evaluation methods; ethics play a negligible role. The MEANS collection has tremendous influence on quality discussions in the evaluation of EU-financed programmes, particularly in countries lacking their own evaluation standards. European countries are very familiar with the requirements found in the MEANS handbooks ⁽⁴³⁾. Some national governments, such as the Finnish, have adopted these criteria (Uusikyla and Virtanen, 2000). The EU Commission also uses the MEANS criteria to assess evaluation reports, grading them from one to four ⁽⁴⁴⁾. Because the MEANS criteria are also implemented for intermediary reports, they can acquire the character of minimum standards, although they also consider unforeseen circumstances. Since the MEANS criteria are similarly concretised in the DeGEval, SEVAL and JC standards, collaboration on further development could be beneficial.

The International Labour Office (ILO) has devised guidelines for the external evaluation of its own programmes, including vocational training. The ILO is active in many developing countries, as well as in new European States such as Poland, and its guidelines apply to Europe. They specify evaluation aspects which should be given the most consideration: effectiveness, relevance, efficiency, sustainability, causality, unexpected effects, alternative strategies and specific ILO concerns. They regard stakeholder involvement and the role of evaluators to be particularly important aspects of approaches to independent evaluations. From an organisational perspective, they focus on composition, schedules and information sources. The topics 'qualification profiles and responsibilities of external evaluators' and 'role of the stakeholders' are addressed in further sections. To summarise, the ILO guidelines are much more concrete and detailed than the DeGEval standards, which are,

however, enhanced by the highly extensive and comprehensive material in the JC standards.

The Public Management Service's *Best Practice Guidelines for Evaluation* are intended to help OECD Member States improve the utilisation of evaluations in performance management systems. They primarily consult those people responsible for the political control of evaluations (governmental organisations, politicians and leading public servants). There are a total of nine guidelines, each with two to five itemised paragraphs listing recommendations. They impose strong demands for involvement of stakeholders. The development of an evaluation culture is also seen as an important task at the level of supranational organisations. The PUMA guidelines share many features with the 'standards' and give contributing input for decision-making a far higher priority than other objectives. They create distinct tension between the decision-maker approach and the participatory one.

No empirically supported statements can be made on the scope and depth of the application of standards in Europe. A few prominent examples are known to the authors.

Switzerland has a leading position in Europe with its far-reaching evaluation culture and the use of evaluation standards. The work of Widmer has created a relatively dense information base. In a recent publication (Widmer, 2003) he lists six meta-evaluations (five from Switzerland) which used the JC or SEVAL standards to assess several (in one case, 43) evaluations. However, none of the three comparative case analyses Widmer conducted himself, covering a total of 18 evaluations, deals directly with VET programmes.

German VET evaluations use JC standards in isolated cases (Peltzer, 2002). Further examples of the application of JC standards have been found in Europe outside the VET context (e.g. Jacob and Varone, 2002). However, they have been consulted relatively rarely in Spain, although a Spanish translation is available. In Spain they are also seldom employed for meta-evaluations (Bustelo Ruesta, 1998). In the US, where the JC standards have been established longest, hardly any publications exist on

⁽⁴³⁾ They are frequently included in meta-evaluations, although often remarkably cursorily. One exception is Polverari and Fitzgerald, 2000; p. 30. These meta-evaluations also often refer to the JC standards, although again usually without explicitly and specifically citing individual standards.

⁽⁴⁴⁾ Stated by European Commission employees at the European Evaluation Society Conference (EES, 2002).

systematic surveys on the adoption and application of the standards ⁽⁴⁵⁾.

In conclusion, we can say that there are no serious discrepancies or contradictions between the European evaluation norms presented here and the DeGEval standards. The various sets of standards simply have separate focuses and are concretised differently. Some are formulated generally, others contain more precisely defined rules. Many of the standards discussed above correspond to central elements of the DeGEval standards. This makes them a suitable specialist evaluation reference, along with the JC standards.

3.3. Transferability of evaluation standards to VET

The US standards were originally developed for the educational sector. Most of the examples in the JC handbook come from elementary schools, high schools, colleges and universities, but also from vocational training and social work. Perusal of the terms and definitions of the 30 individual standards reveals that only one contains educational terminology. This is the standard JC-P1 (service orientation support) which demands that evaluations 'help ensure that educational and socialisation objectives are appropriate' ⁽⁴⁶⁾. Because of its specific nature this standard is not included in the German and Swiss standards, which are designed to be general and applicable to all evaluand fields.

The handbook (JC, 1994) illustrates each JC standard with positive and negative examples and their analyses to clarify the actual text and detailed guidelines. Seven of these examples involve case studies from in-company vocational training and are therefore directly applicable to VET (Widmer and Beywl, 2000, p. 249).

- (a) In the illustration contained in JC standard U5 (report clarity) a vocational training planning team commissions an evaluation of a training

programme and expects a written report with suggestions for improvement. The reporting could have been better.

- (b) JC standard U7 (evaluation impact) gives the example of a formative evaluation of performance-based training in the industrial sector. A checklist devised by trainers and evaluators helps record the (altered) behaviour of trainees. The overall design of this evaluation was excellent. The fact that all stakeholders remained motivated until the very end is a major success.
- (c) JC standard P2 (formal agreement) is illustrated by a case where the staff training manager of an enterprise has sought the advice of an evaluation consultant. The consultant has deviated from the agreed evaluation plan. She has conducted a written survey of graduates of a management training course, something that was not originally stipulated. In this case the contract between the commissioner and the evaluator should have been updated.
- (d) JC standard P4 (human interactions) contains the following illustration: an internal evaluator is to collect information on the training needs of secretaries in all units of the company, in order to test the effectiveness of the current programme and propose changes. She

Table 4: **VET examples in the JC standards (1994/2000)**

No	Standard term
JC-U5	Report clarity
JC-U7	Evaluation impact
JC-P2	Formal agreement
JC-P4	Human interactions
JC-A1	Programme documentation
JC-A2	Context analysis
JC-A12	Meta-evaluation
Source: JC, 1994	

⁽⁴⁵⁾ A panel discussion on *Applying program evaluation standards* took place at the American Evaluation Association annual meeting sessions in November 2001. In her paper *Standards-based processes for program evaluation at SERVE*, Mary Sue Hamann presented several SERVE guidelines. SERVE is an educational organisation serving six American States. The four binding guidelines on evaluation bids, evaluation contracts, evaluation designs and evaluation reports are based on the JC standards. In his paper on *Making use of evaluations standards routine* Ken Town from the University of Southern Maine described a long-term initiative of the Institute for Public Sector Innovation (IPSI). The intention is systematically to improve evaluation competence among IPSI employees and in the organisation as a whole. The initiative is based on the JC standards. Most of the employees are not evaluation experts so the multidisciplinary JC standards are a useful resource.

⁽⁴⁶⁾ We advocate restoring a standard of this nature to VET evaluations (Section 6.5).

conducts focus group interviews but these antagonise a leading personnel manager and have to be abandoned. Forming an advisory committee for stakeholders at the start of the project would have helped avoid this problem.

- (e) JC standard A1 (program documentation) is illustrated with an evaluation of an in-company technical training programme including computer-based training. The members of the supervisory panel watch a demonstration of how computer-based training works. This was the only way to acquire the knowledge needed to tackle the core questions of the evaluation.
- (f) JC standard A2 (context analysis) is explained through an evaluation of the effectiveness of a training programme for sales representatives. The interviewees were members of various company departments, and the findings of several successive focus group surveys were very different. The evaluator was initially unaware of the personnel changes in the departments which had produced the inconsistent assessments. This important contextual information was required at the planning stage.
- (g) The following case exemplifies JC standard A12 (meta-evaluation). An organisation wants to initiate a series of follow-up evaluations to improve its training courses. A meta-evaluation revealed that most of these follow-up evaluations were not completed. The meta-evaluation inspired new impetuses for a future evaluation system.

This brief overview of the JC standards relevant to VET demonstrates that they all have a fundamental similarity. All the examples portray evaluations of individual education and training programmes within enterprises. However, in VET, evaluations of larger programme systems are just as relevant as, for example, evaluations of government initiatives or VET subsystems, or of EU-wide support programmes. The VET spectrum is manifestly broader than the examples from the US handbook suggest (cf. this paper's Outlook, Section 7.3). The following chapter of this paper will examine what other VET requirements need to be addressed.

One of the standards' fundamental tenets is that they can be applied to a broad spectrum of political fields. Stockdill (1986) interviewed experts to investigate whether the JC standards for evaluation were appropriate for the US business world. He established that the standards were also suitable for personnel development and the evaluation of other human resource development tasks in the profit-making sector. The original US standards began in the field of education and were then applied to programme evaluations in other policy areas in Europe. The diversification evidently influenced the DeGEval and SEVAL adaptations.

Since the evaluation standards originated in education and are meant to be applicable to all policy areas, we must assume that they are also valid for VET ⁽⁴⁷⁾. We do not consider it necessary to alter the names and texts of the general evaluation standards. We feel that the explanations, and particularly the illustrative examples, which illustrate and discuss good and bad applications of standards would be particularly beneficial, making standards much more accessible to VET specialists (recommendation 6).

An important connection exists between evaluation standards and programme standards. Evaluators must systematically develop scientific findings and academic theories, fundamentals of the evaluand field and demands for quality and harness them in the planning of evaluations, to respond to the requirements of standards G1/A1 (description of the evaluand) and G2/A2 (context analysis). We also suggest inserting an additional evaluation standard specifically for VET to support the utilisation of scientifically founded, specialist or professional programme standards in VET evaluations.

This standard, Quality orientation support in vocational training, could read as follows: 'Evaluations should assist VET policy-makers and programme managers to meet quality requirements within the vocational training sector (VET standards). These particularly include standards which require evaluations to consider the needs of target groups, social partners and society, have a scientifically founded theoretical and teaching concept, help shape the structure and

⁽⁴⁷⁾ A current textbook on continuing training evaluation (Reischmann, 2003) features the DeGEval standards, explicitly validating their use in VET.

organisation of political education and help manage educational processes and ensure the profitability of VET activities.’ The explanatory notes on this standard should mention well-known, recognised VET standards and point the way to the most relevant sources.

An additional note to JC standard P1 (service orientation) should mention that evaluations are

meant to support decision-makers, sponsors and programme managers in tailoring their VET policies and programmes to the needs and situations of the target groups, and to promote gender mainstreaming and social inclusion (recommendation 8).

4. Dialogue with German and Austrian VET experts on evaluation standards

The following chapter will first describe how the discussions were implemented in the two countries. The third section will summarise the results of the debates as hypotheses.

4.1. Focused events with the German Federal Institute for Vocational Training (BIBB)

The BIBB ⁽⁴⁸⁾ coordinator for additional qualifications, learning organisations and process orientation organised a series of workshops from autumn 2000 on concomitant-research methods. The sessions focus on research support ⁽⁴⁹⁾ for pilot projects on learning organisations, additional qualifications, process-oriented vocational training and cooperation between learning locations. The pilot projects commissioned by BIBB test the practicality of innovative developments in initial and continuing vocational training. The aim is to translate their findings into vocational training practice ⁽⁵⁰⁾. Their dual purpose is to improve the areas of vocational training practice covered and, at the same time, gain insight into the evaluand's field.

During the second workshop in April 2001, the topic of standards for evaluation was introduced in a lecture ⁽⁵¹⁾. The short discussion focused on the conflict of roles consulting researchers face from the various expectations of stakeholders such as

administrators, practitioners and academics. How can close contact between researchers – indispensable for the application of findings in learning organisations – be guaranteed while still ensuring that researchers remain impartial and independent ⁽⁵²⁾?

Because of their significance for the further exchange of experiences, standards for evaluation, adopted in the interim by the DeGEval, were the main topic of the third meeting, held in Frankfurt am Main in October 2001. The keynote was Prof. Klaus Jenewein's ⁽⁵³⁾ detailed meta-evaluation lecture relating to a pilot project, Development of occupational skills via a contract type concept for initial vocational training. To test the standards, he applied them to his concomitant research and performed a meta-evaluation of his completed research. Jenewein concluded that most DeGEval standards can be applied to evaluations of vocational training pilot projects. He claimed that one problem was the plethora of objectives which concomitant research into pilot projects must pursue (including development, summary assessment, promotion of mainstreaming, legitimisation), particularly concerning the demands the standards make for impartiality and propriety in testing (F4/P4 and F3/P3). He maintains that, since pilot projects test innovative ideas and vocational training content, rigid demands for valid and reliable data collection and assessment often cannot be met (G5/A5 and G7/A7). He also doubts it is possible to measure

⁽⁴⁸⁾ BIBB was founded in 1970 pursuant to the Vocational Education and Training Act of 1981. The federal public law body is supported by the federal budget. It investigates initial and continuing vocational training practice in enterprises. This involves testing new approaches to initial and continuing vocational training and, in conjunction with the social partners, setting company regulations on vocational training and career advancements.

⁽⁴⁹⁾ Approximately 20 people participate in the workshops. They are usually concomitant researchers with many years experience in initial and continuing vocational training.

⁽⁵⁰⁾ The institute is legally obliged to promote pilot projects and their supporting research. This is specified as an objective in its work programme.

⁽⁵¹⁾ Wolfgang Beywl, talk and transparency presentation on *Evaluationen an lernende Organisationen anschlussfähig machen – Hinweise und Standards für Programmevaluationen aus der Evaluationspraxis* (Making evaluations of learning organisations compatible: instructions and standards for programme evaluation resulting from practice).

⁽⁵²⁾ Dorothea Schemme, minutes of the second session of the concomitant-research-methods workshop on 26 April 2001 in Stuttgart.

⁽⁵³⁾ Prof. Klaus Jenewein works in the Vocational Education and Technical Didactics department of the University of Karlsruhe, Germany.

the cost-benefit ratio, required to establish the efficiency of the evaluation (D3/F3) and believes the problem is aggravated by, or even conflicts with, the basic values (ultimately criteria) of the various stakeholders (e.g. target groups, sponsors, companies, schools).

The proceedings of the third workshop emphasise the analytical distinction between a programme and its evaluation, stated in the introduction to the DeGEval standards. This dichotomy provides the opportunity to specify the role of concomitant researchers and can help improve transparency and awareness of the dilemmas outlined by Jenewein. We can define roles and requirements for the interaction between programme managers and evaluation managers, making the performance of both more verifiable and controllable. Given the dual purpose of pilot projects – to improve practice and gain insight – ‘pragmatic orientation, trans-disciplinary procedures and a reduction in applied research usually have priority over the precise construction of perfect scientific use of tools known from basic research into single disciplines’⁽⁵⁴⁾.

In May 2002 BIBB scheduled an internal colloquium, open to its entire staff, on *Evaluation standards and their application in vocational training*. The aim was to introduce dialogue on the adaptation of the DeGEval evaluation standards for vocational training as implemented by the BIBB itself or subcontracted. Participants were to air questions on the standards, establish further discourse requirements and discuss how the dialogue should be continued.

Some 30 BIBB employees from many different initial and continuing vocational training fields took part in the two-hour event. Most participants work primarily or partly as consultants or evaluators. After an introduction to the DeGEval standards⁽⁵⁵⁾, the discussion focused on the following aspects:

- (a) relations between quality management and evaluation/potential for synergy;
- (b) differences and overlaps of concomitant research and evaluation;
- (c) validity of the standards for self-evaluations;

- (d) suitability of the standards for comparative evaluations;
- (e) suitability of the standards for meta-evaluations;
- (f) fears that application of the standards might tie down too many resources;
- (g) warnings about (potential) contradictions between individual standards;
- (h) lack of guidelines, frequent errors and illustrative examples in the JC unabridged version;
- (i) intercultural transferability of standards originating in the US.

Points (a), (b) and (c) concern the boundary between evaluation and other forms of academic support for, and assessment of, programmes and projects in vocational training. Such support is the concomitant research approach commonly practised in vocational training, although it is methodologically less elaborate than evaluation, since specialist textbooks are rare. However, quality management or quality assurance, or even systematic development and testing of quality along the lines of consumer reports, are certainly of interest to vocational training. After all, everyday language equates the self-evaluation approach with self-assessment, although in Germany the former has been much more sharply defined and presented in several monographs.

Points (d) and (e) concern the scope of validity of the standards for comparative evaluation and meta-evaluation of programmes or projects. The fact that this was questioned makes it clear that the DeGEval standards – particularly the terse 25 individual standards – are not self-explanatory. We recommend always consulting explanations and the additional US sources as a supplementary reference.

Points (f), (g) and (h) cover queries and critical observations on the evaluation standards. It is clear that the evaluation standards (or the codes of ethics) present fundamental dilemmas. Professional debate and a feedback process are necessary to ensure that they are reliably put into practice. On the one hand, evaluators feel that over-demanding or operationalised standards or a high density of rules might overtax practical evaluations. The standards explicitly refer to this

⁽⁵⁴⁾ Dorothea Schemme (BIBB), minutes of the third meeting of the concomitant research methods workshop of 12 February 2002.

⁽⁵⁵⁾ The introduction followed this basic structure: a brief definition of ‘evaluation’, the origins and system of the standards, an example of the implementation of a standard.

danger, particularly in the individual standard D1/F1 (appropriate procedures). On the other hand, workshop participants stressed that applying the individual standards to practical evaluation projects could lead to contradictory demands. In such cases compromises will have to be reached and priorities assigned to 'competing' standards. This is another problem which the DeGEval and JC standards mention distinctly. Finally, BIBB specialists would like more concrete and more palpable standards to give evaluators, in particular, as much tangible help as possible and to tailor the standards for use as a (self-)education programme.

Point (i) amounted to a brief expression of the general concern about the transferability of standards from a different culture and society.

4.2. Focused meeting with the Austrian Federal Institute for Adult Education (BifEb)

Strobl am Wolfgangsee in Austria hosted a three-day colloquium of around 14 hours from 2 to 4 April 2002. Its title was *Quality development in adult education: evaluation standards and methods*. The Federal Institute for Adult Education (BifEb) organised the event ⁽⁵⁶⁾.

Approximately a quarter of the 17 participating specialists work primarily in the vocational training field. Most are trainers who devise or conduct continuing training courses themselves. A few are external evaluators of general or in-company continuing training. The participants had little or no prior knowledge of the standards. They familiarised themselves with the system and content of the standards through lectures, individual and partner activities on the text of the DeGEval standards, and application to their own, usually internal, evaluands. Their primary concern

was to put the standards into practice when planning their own evaluations and commissioning them. The course focused on evaluation control, data collection and interpretation and reporting and utilisation of findings. Two tools were employed to assess evaluation standard suitability in vocational education and continuing training:

- (a) a poster survey on application of the standards; towards the end of the seminar, participants were asked to note their responses to the following questions on big posters:
 - (i) what consequences do you think the standards should have for your work?
 - (ii) what steps should managers of continuing training institutions take with regard to the standards?
 - (iii) how should adult education and continuing training legislation and public sponsors react to the standards?
 - (iv) how should DeGEval address standards in the area of adult education and continuing training?

All 17 participants contributed to the poster survey ⁽⁵⁷⁾. Their comments were subsequently discussed in a plenary session, which made it possible to acquire a deeper understanding of some points and to ascertain the intention of each remark;

- (b) short printed questionnaires on the suitability of the DeGEval standards; questionnaires were distributed to gain insight into how suitable the participants thought the DeGEval standards were for adult education and vocational training. They addressed the following topics ⁽⁵⁸⁾:
 - (i) arguments for the three evaluation purposes (preparation for decision-making, improvement and gaining insight);
 - (ii) distinction between formative and summative activities;
 - (iii) interpretation of the standards as maximum standards;

⁽⁵⁶⁾ BifEb was founded in 1956 and is the training institute for adult education, supported by the Austrian Federal Ministry of Education, Science and Culture (according to Article 11, Paragraph 1 of the Adult Education Promotion Act of 1973). It employs 30 members of staff with and without educational qualifications. It targets multipliers inside and outside traditional adult education. Its main focuses are vocational and further training of staff, training management, training consultancy, programme creation, organisation, supervision, evaluation and new approaches to teaching and learning. Available from Internet: <http://www.bifeb.at> [Cited 29.10.2003].

⁽⁵⁷⁾ The responses were incorporated into the findings of the event.

⁽⁵⁸⁾ The topics were chosen on the basis of the questions posed during the CFT 13 subproject and complemented by points of the discussion during the first working group meeting on the third Cedefop report on vocational training research in Europe, 28 February to 1 March 2002, Thessaloniki.

- (iv) unsuitable individual standards;
- (v) missing individual standards;
- (vi) suitability of terminology;
- (vii) limits of evaluation;
- (viii) European dimension;
- (ix) standards revision processes.

Seven participants completed and returned the forms ⁽⁵⁹⁾. The comments made it clear that it was difficult for the participants to answer very specific questions. This was particularly the case for questions concerning unsuitable individual standards, missing individual standards and the European dimension ⁽⁶⁰⁾.

4.3. Conclusions from the dialogues

Neither the Austrian nor the German experts saw any fundamental restrictions to the application of the DeGEval standards to the field of initial and continuing vocational training. Members of the widely varying academic cultures involved in VET evaluations and research regard certain individual standards as unfathomable, insufficiently defined, vague, contradictory and possibly irrelevant. However, they acknowledge that the standards, and the theories and experience they embody, offer tremendous learning and development potential for evaluations and impact investigation in initial and continuing vocational training.

There is accepted applicability in vocational training. No doubts were expressed on the standards' transferability to vocational training as an evaluation with specific institutional arrangements (for example, the dual system of vocational training in Germany). The experts do not propose specific adaptation, although they would like to see certain standards illustrated through concretely demonstrated examples from initial and continuing vocational training.

There is uncertainty as to validity for different academic cultures. Evaluators (researchers) who

have been working for many years within a particular discipline or theoretical tradition have initial concerns that their methodology might not be adequately covered by the evaluation standards. The researchers working for and collaborating with BIBB felt this way. We assume that representatives of other schools may also not initially consider how the standards might apply to their preferred approach. For example, some researchers consider the use of experimental and quasi-experimental design the litmus test of the quality of evaluations ⁽⁶¹⁾.

There is ambivalence with regard to maximum standards. The conception of the DeGEval standards as 'maximum standards' with a primarily orienting function, intended to inspire dialogue on the quality of evaluations, was received ambivalently. Some ascertained (although maybe hesitantly) one advantage of maximum standards to be that they can refer to many different approaches and types of evaluations and impact investigations. However, the representatives of certain 'schools' regretted the lack of the prescription of obligatory minimum standards. Concomitant researchers, for example, would want the project evaluators to be vocational training specialists and perhaps to have conducted their own independent research based on vocational training theories, or to have published articles in this field. However, advocates of more experimental approaches would desire minimum requirements including such features as control group designs, or specific mandatory procedures for random selections.

There are concerns that links to evaluation theory are not sufficiently explicit. The notes explaining the standards state that there are 'numerous different approaches to professional evaluation' and that these vary markedly depending on epistemological approach, discipline and professional ethics. The pluralistic foundation of the standards is not immediately clear to experts the first time they read them. They

⁽⁵⁹⁾ The results have been incorporated into the proposals.

⁽⁶⁰⁾ No question was posed as to whether the standards are equally applicable to evaluations in the micro-, meso- and macro-areas, since the experiences of most of the participants in Strobl have mainly been in the micro-area (organising learning processes, establishing curricula), rarely in the meso-area (evaluations of [external] company continuing training systems) and not at all in the macro-area (vocational training policies and their effects on the whole of society and the general economy). That also explains the difficulty they had expressing an opinion on the European dimension.

⁽⁶¹⁾ We received a refusal for the expert surveys discussed in Section 5. The reason given was that our questionnaire did not include explicitly the relevance of (quasi)experimental designing.

often worry that the standards will have a restrictive effect on the approach they advocate, or even exclude it entirely. Besides, the various functions of evaluation developed by the newer evaluation theorists (e.g. proactive, clarifying, interactive, monitoring and impact evaluation [Owen and Rogers, 1999]) and Stufflebeam's typology with around 20 evaluation models (2001) are not sufficiently recognised as linked to the pluralistic function of the evaluation standards (recommendations 3 and 7).

There is a perceived conflict of roles in terms of utility, accuracy and independence. University academics in particular, but also those at public and private research institutes, feel that they face a strong conflict of interests. The standards, and their four main tenets of utility, feasibility, propriety and accuracy, have increased this awareness. When public or private bodies commission evaluations and impact analyses, they usually expect immediately utilisable findings. Sponsors and heads of facilities where the data is collected prefer streamlined procedures and tools which do not disrupt current initial and continuing vocational training. Data protection regulations also impose some major limitations, particularly when the performance of teaching personnel is directly or indirectly described or judged by evaluation processes⁽⁶²⁾. Furthermore, the accuracy group imposes strict requirements of empirical social and economic research on, for example, the validity of tools and the reliability of data collection. Descriptions and assessments must be independent. All these demands coming from different quarters may compete with one another in evaluation practice, and situations could arise where they cannot be reconciled. The evaluation standards expose these contradictions but do not propose any general solutions.

There are limitations to the possibility of self-evaluation. In Germany, the concept of self-evaluation has been widely propagated in the social services and school system, partly by several monographs and manuals⁽⁶³⁾. The situa-

tion in Austria and Switzerland is similar. Non-school initial and continuing vocational training programmes have also started to introduce it. The participants in the Austrian seminar, who primarily work as evaluators in small and medium-sized continuing training institutions, expressed particular interest in the self-evaluation approach. Teachers can implement it at the micro-level of teaching and learning processes. It has few extra costs, e.g. for external evaluation consultancy. Well-versed professional experts are initially uncertain whether the evaluation standards also apply to self-evaluation. The fact that they do not, and that DeGEval has developed separate self-evaluation standards, becomes clear from the explanatory notes, but has often been the subject of inquiries. We should ask whether other European countries are familiar with, and use, self-evaluation approaches to VET evaluation or whether they rely solely on external or internal independent evaluation⁽⁶⁴⁾.

The brief summary of standards is perceived to have limited usefulness. Interested parties often read only the brief summary of standards, which is approximately three pages long. Explanations of the individual SEVAL standards cover about one page each. The DeGEval standards are accompanied by similar explanations by the Standards Commission, but these are not a formal component. The JC standards contain several more markedly operationalised guidelines in addition to the explanatory notes for each standard. The committee recommends compliance with these guidelines. They include a list of frequent errors and a few annotated examples, which help elucidate the applicability of a certain standard. During the workshops several people expressed the desire for more comprehensively annotated specifications similar to the JC publication. If possible, these should be supported by illustrative examples of evaluations in the field of initial and continuing vocational training.

The clarifying function of the standards was positively received. Many respondents praised

⁽⁶²⁾ The DeGEval standards, like the JC and SEVAL standards, thus emphasise that they are not suitable for personnel evaluations. For that purpose the JC published the *Personnel Evaluations Standards* as early as 1984. It did not publish its *Student Evaluation Standards* until 2002.

⁽⁶³⁾ It has major similarities to the concepts of empowerment and collaborative evaluation. If, and when, evaluation specialists are involved long term in these projects, they undertake a role as teachers and facilitators (Fetterman, 2000).

⁽⁶⁴⁾ DeGEval's Social Services working group has its own set of standards specifically tailored to self-evaluation. Available from Internet: http://www.degeval.de/ak_soz/index.htm [Cited 30.10.2003].

the fact that the explanatory notes on the DeGEval standards defined terms. These include, for example, the difference between stakeholders, addressees and users. There is an analytical distinction between the purpose of the evaluation (and the evaluation approach) on the one hand and the aims of the programme/eval-

uand (and its approach) on the other. This facilitates making an analytical division between the role of evaluation and responsibility for the programme, particularly during formative evaluations or concomitant research. These valued aspects of the standards underline the importance of annotated explanations.

5. E-mail survey of evaluation experts in Europe

A further element of the study into evaluation standards was an e-mail survey of expert opinions. The poll addressed quality requirements for evaluations in VET. First of all, this chapter outlines the questions and the process; the sampling method is also described. The answers to questions 5 to 7 provide an overview of the experts' attitudes towards evaluation standards, their preferences and their familiarity with various sets of standards. Questions 8 to 12 asked for critical comments on the advantages and disadvantages of the standards and on the establishment of basic values which the evaluation quality requirements should contain ⁽⁶⁵⁾.

The survey yielded answers to the following two central questions:

- (a) does Europe need a codified rule book in the guise of evaluation standards to ensure and increase the quality of VET evaluations?
- (b) what cultural and professional values and requirements should be addressed in such a code?

Experts on evaluation and/or vocational training from various European countries were approached by e-mail. We had had no prior contact with most of these experts. Most of them were located through the support of national evaluation associations and members of the board of the European Evaluation Society. We also utilised ERO-CALL, a mailing list mainly featuring VET experts, to invite people to participate in the survey. The questionnaire was sent as a text file. We asked respondents to recommend other experts for the survey. We subsequently contacted them.

The three-page questionnaire is written in English and consists of a total of 15 items. Seven

are closed questions (one with several sub-questions). The eight open questions gave respondents the opportunity to state their opinions and provide feedback.

The e-mail questionnaires sent to the experts were accompanied with the request to fill them out electronically and return them by e-mail or to fill them out by hand and fax them. We chose to conduct the survey by e-mail since most initial contacts had been made via this medium and because it accelerated the procedure. The questionnaires went out late in August 2002 and the deadline for their return was 4 October 2002.

We used SPSS to process the quantitative data. We analysed the content of the qualitative data on the open questions. Questions 13 to 15 were merely devised to assist organisation of the study ⁽⁶⁶⁾ so these answers do not feature in this report. The following tables include the text of the original questionnaire for clarity's sake.

5.1. Profession and nationality of respondents

Limited resources restricted the survey to a small sample from the outset, so it cannot claim to be representative. 19 of the 30 experts who received a questionnaire replied ⁽⁶⁷⁾. This is a satisfactory response rate ⁽⁶⁸⁾. The total of 19 returned questionnaires can be seen as a pool of trends and indications that can be scrutinised in conjunction with other investigations to make valid interpretations.

⁽⁶⁵⁾ See the questionnaire in Annex 2

⁽⁶⁶⁾ Contact addresses, other contact recommendations, hints on relevant literature.

⁽⁶⁷⁾ Please see the list in the Annex 2.

⁽⁶⁸⁾ Regrettably, only one person from the UK responded by the deadline (Figure 3).

Table 5: **Primary position in evaluation**

Your primary position in/to Evaluation	Frequency	Percentage
Evaluator	9	47.4
Client/sponsor/commissioner	3	15.8
Programme director/programme staff	1	5.3
Other	6	31.6
Total	19	100.0
Source: author's representation		

Around half the respondents were evaluators, while three commissioned or sponsored evaluations. One person was a programme manager or a member of programme staff. Six respondents

had posts outside evaluation. One was an evaluation handbook author, four were researchers and one a regional administrator dealing with evaluations.

Table 6: **Respondents' professional background**

What is your main professional background	Frequency	Percentage
Economics	3	15.8
Social and political sciences	12	63.2
Liberal arts including pedagogic	4	21.1
Total	19	100.0
Source: author's representation		

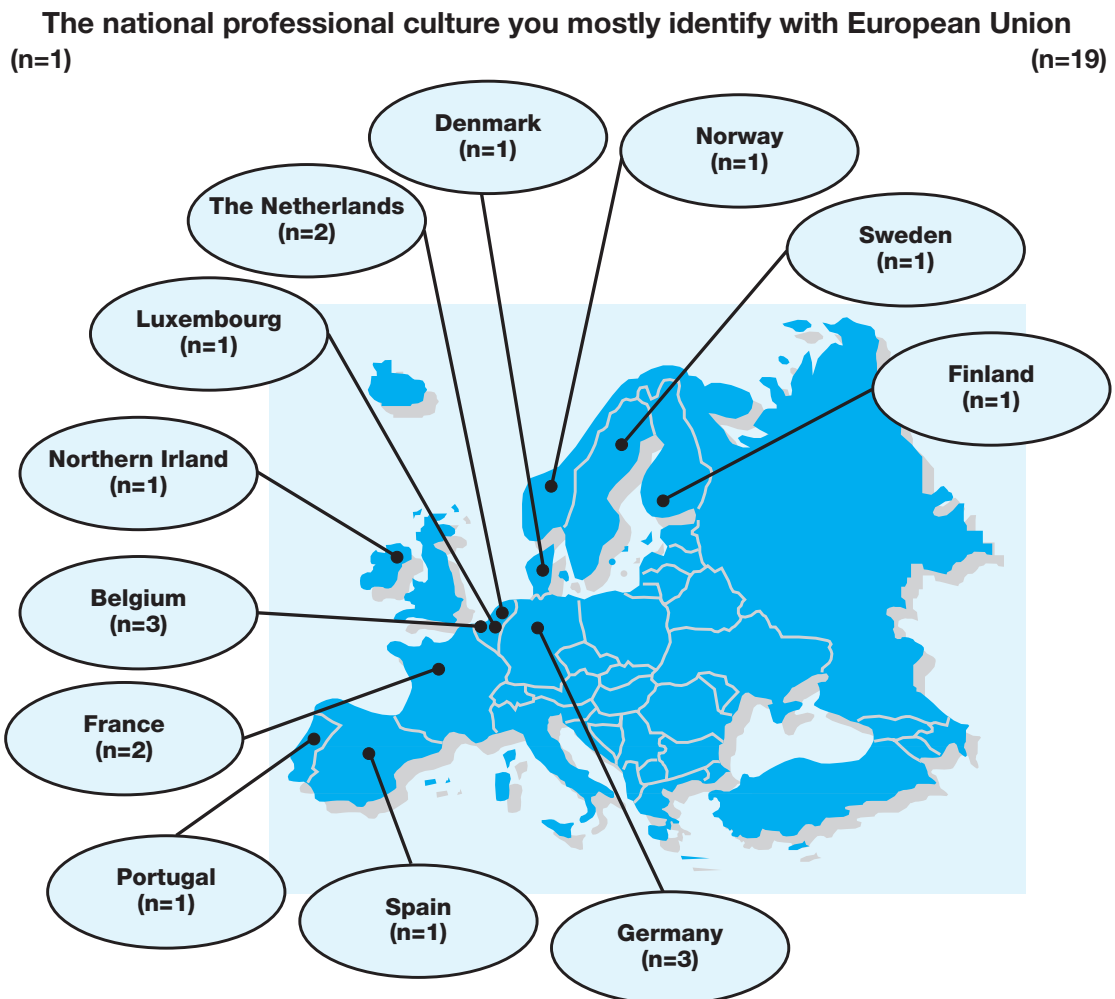
The professional background of around two thirds of the respondents was social and political sciences. Four respondents represented the liberal arts (including teaching) and three were economists. Two respondents also entered natural sciences as a secondary field. No engineers participated. Seven of the nine evaluators were social and political scientists.

Three respondents each identified themselves with the German and Belgian professional cultures. France and the Netherlands were each named twice. One respondent each cited Denmark, Finland, Sweden, Norway, Northern Ireland, Luxembourg, Portugal and Spain. One respondent named the culture of the EU.

Table 7: **Respondents' relation to VET**

What is your relation to VET?	Frequency	Percentage
VET is my main/most relevant working field	3	15.8
VET is one of my most relevant working fields	6	31.6
VET is a known field for me but I am (nearly) not active in	10	52.6
Total	19	100.0
Source: author's representation		

Figure 3: Respondents' identification with national professional cultures



Source: author's representation

Over half (10 respondents) of the international experts said that they were familiar with VET but that they were rarely, if at all, involved in it. Around a third (six) identified VET as a relevant working field for them. Only three respondents named VET as their main or most relevant working field. This distribution suggests that evaluation experts in the field of VET who also have a thorough knowledge of standards or other evaluation quality norms hardly seem to exist or are difficult to identify.

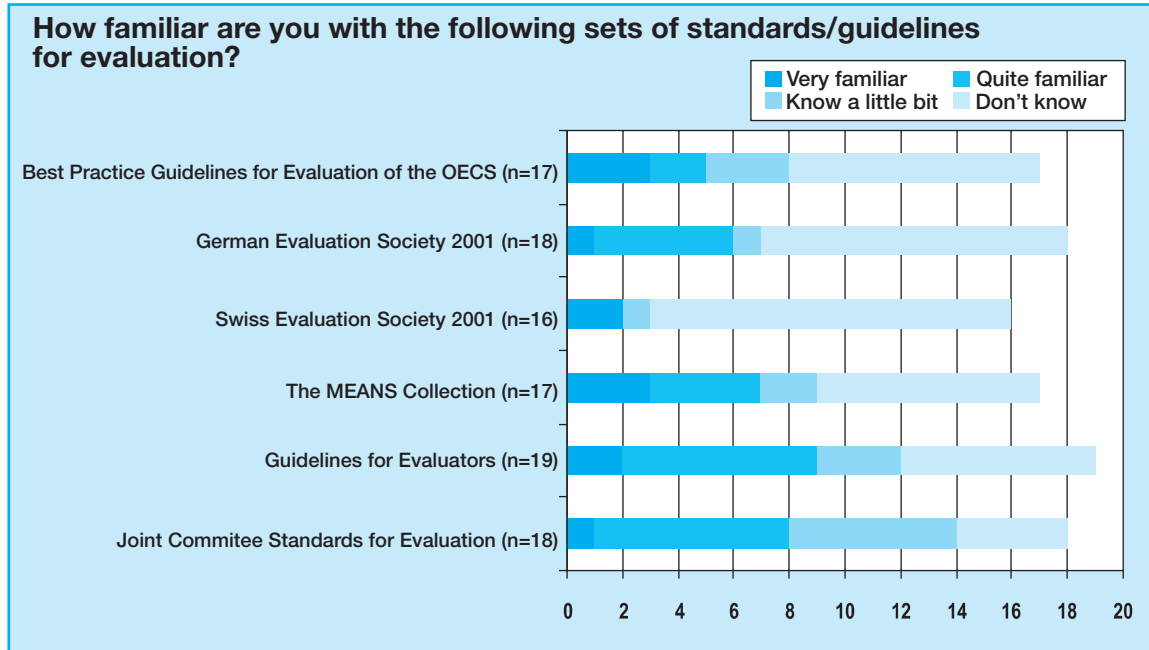
Of the nine respondents who named VET as their main or at least one of their most relevant

working fields, five are evaluators. Of the remaining 10 for whom VET is not a central field 4 are evaluators.

5.2. Assessment of existing evaluation standards

The first block of questions covers the degree of familiarity with various guidelines for evaluation and attitudes to evaluation standards in general as well as to the two alternatives, minimum standards and maximum standards.

Figure 4: Respondents' familiarity with various sets of evaluation guidelines



Source: author's representation

The respondents reported on how familiar they are with various current standards or guidelines for planning and implementing evaluations. The examples given were the *US joint committee standards for evaluation* and the *Guidelines for evaluators* published by the American Evaluation Association, the European Commission's *MEANS Collection*, the SEVAL standards, the DeGEval standards and the OECD's *Best practice guidelines for evaluation*. Respondents could also name other standards or guidelines.

A total of 15 respondents identified at least one standard with which they were familiar. The best known were the *Joint Committee Standards* and the *Guidelines for Evaluators*. Twelve of the respondents were familiar with the former and 14 with the latter, to at least some degree. They were followed by the *MEANS Collection*, OECD's *Best practice guidelines* and the DeGEval standards.

Four experts mentioned one other set of guidelines, a theory or literature that they consult,

and two experts named two publications. These were quoted as:

- (a) the Finnish Evaluation Society's Ethics of evaluation;
- (b) French system AFD, IEPF rules;
- (c) Investors in people standard, United Kingdom;
- (d) our own framework for evaluation, including distributional effects, concepts borrowed from A. Sen, duration analysis, cost-benefit analysis;
- (e) ISO quality measurement more than standards;
- (f) range of textbooks on evaluation theory.

These answers provided interesting insights into what additional sources could be considered in the further development of evaluation standards in Europe.

In general we discovered that no matter which standards or guidelines we listed, a maximum approaching half the respondents were familiar with them to a considerable or certain degree. The majority of the respondents knew most of the standards listed only fleetingly or not at all. The JC Standards clearly are best known. Only two respondents had not heard of them.

Table 8: **General assessment of evaluation standards**

General position to standards for evaluation	Frequency	Percentage	Cumulative percentage
Standards are absolutely necessary	7	36.8	36.8
Standards are important	6	31.6	68.4
Standards could be useful	6	31.6	100.0
Standards for evaluation do not matter	0	0.0	100.0
Standards for evaluation are not necessary or even harmful	0	0.0	100.0
Total	19	100.0	
Source: author's representation			

All the experts had a positive attitude towards standards for evaluation. Over a third feel that standards are absolutely necessary. A third believe that standards are important and the remaining third think that standards could be useful but do not yet seem to be sure whether they actually will be. None of the respondents ticked the fourth or fifth option, that the standards do not matter or are unnecessary or even harmful ⁽⁶⁹⁾.

On the open question 8 nearly all respondents argued for the intensive use of standards in VET evaluation ⁽⁷⁰⁾. Many emphasised that standards lead to improvement in the quality of evaluations. Cited advantages are higher professionalism, the possible use of the standards as a study aid and means of establishing uniform terminology, improvement in the utility and significance of evaluation projects, and, especially, improved transparency and comparability of evaluation projects.

Quoted responses include:

- (a) to diminish biases in evaluation. To get more justice for everybody who is evaluated;
- (b) I was active in a Dutch consulting project on examination and evaluation in VET. I discovered the importance of a minimum language to be able to exchange between the various educational tracks;

(c) I think that standards utilisation is the best way to increase the quality of evaluation by the development of a common framework for all the stakeholders (commissioners, evaluators, [...]);

(d) in countries where evaluation is just being introduced evaluation can have very important functions in changing organisational cultures and the functioning of organisations in many ways. There should, however, be clear standards in order to protect all participants.

The respondents also see in the standards an improved opportunity for evaluators' work to appear more legitimate, transparent and verifiable to outsiders. This could help protect all stakeholders:

- (a) raise credibility and professionalism of evaluation and evaluators, provide a valuable checklist for evaluators and those wishing to appoint evaluators, identify expectations and benefits from the evaluation process [...];
- (b) in [our country; W.B./S.S.], in the case of absence of standards, the profession will never exist as a special profession of a group of professionals, whether in VET or any other domain;
- (c) enhance relevance, usefulness, and utilisation of evaluation;

⁽⁶⁹⁾ We cannot exclude the possibility that the eleven people who did not respond to the survey or that other unidentified VET evaluation experts have this sceptical or negative attitude to evaluation standards.

⁽⁷⁰⁾ Only two of the 19 respondents did not advocate more intensive use of standards. One of these was not familiar with any of the standards mentioned in question 7.

- (d) if nothing else, it can at least enhance discussion about the relationship between evaluation and ethics.

Two respondents regretted that evaluations are often understood in a very one-sided manner, being either restricted to their summative function or concentrating purely on short-term effects. They hope that the evaluation standards will help extend the scope and time frame of evaluations.

One respondent stressed that evaluation standards should be generally applicable and not designed specifically for one field, such as VET.

Although many respondents felt that setting rigid standards could jeopardise the plurality and flexibility of evaluation (see the disadvantages mentioned in answer to question 9), several of their colleagues believe that standards can safeguard against one-sidedness and the loss of flexibility by emphasising methodological variety and plurality of perspectives. Observers hope that evaluations based on standards will thrive, since comprehensible guidelines have prepared the ground for reaping tangible benefits. Fewer respondents mentioned the opportunity for professional exchange on the subject of evaluations which the discussion of standards provides; but those who did make mention, value it.

Of the 19 experts who responded to the open question 9, 6 did not identify any disadvantages in a more intensive use of standards ⁽⁷¹⁾. The misgiving most often expressed was that standards could lead to a loss of plurality and flexibility, and thus to rigidity, in the theory and practice of evaluation, creating barriers to innovation. Some respondents believe that the multitude of cultural and historical approaches to evaluation cannot be reflected in standards:

- (a) 'standard' may not always do justice to national/historical idiosyncrasies; an explorative attitude is necessary also in evaluation;
- (b) [...] different evaluation cultures in different countries; lacking flexibility if standards are not further developed/updated; sponsors/donors could feel to be hampered in their programmes;
- (c) there is the problem and fear of harmonisation;
- (d) when something becomes institutionalised and written, many negative, unexpected and unintended side effects may occur, e.g. lip-service kind of talk;

- (e) it is too restraining, the evaluator might lose interesting development features in the field.

An idea expressed almost as frequently was that prescribing standards could lead to mechanical application which would not suit the evaluation focus or the evaluand. Respondents suggested that the alleged objectivity of rigid standards could eclipse the individual (ethical) decisions of the evaluators, and ultimately undermine the real quality of evaluations, if unquestioning obedience to standards were to become the overriding principle:

- (a) [...] they might be applied in a mechanic way if they are too technical. Standards always transport values and methodological as well as theoretical applications that would narrow the scope of approaches and might hinder innovation [...];
- (b) risk to focus the evaluation on the respect of procedure rather than of its purpose. Risk of rough benchmarking and comparison.

The standards could also hinder evaluation. For example, commissioners might employ them as an instrument of control or pressure, or smaller organisations might refrain from evaluating if they are obliged to follow standards slavishly. One person rejected the idea of a possible seal of approval for institutions and/or evaluators. Some warned against competition for distinctions of this nature or recognition for 'compliance to standards':

- (a) if a standard 'kite mark' became attainable it should not prohibit small companies from applying to attain the standard; standards should be reviewed; community development evaluation (which often includes areas of VET);
- (b) it takes a lot of effort by the evaluators.

At this stage respondents also pointed out that standards must be worded very carefully to eliminate the risk of 'poor standards'.

5.3. Further development of evaluation standards

Respondents were asked to decide which of the two following standard types they prefer. Minimum standards are precise, operationally indispensable minimum conditions that the evaluation must fulfil.

⁽⁷¹⁾ Three explicitly answered 'none', three provided no answer.

If one minimum standard is not observed, the evaluation is not acceptable. Maximum standards describe desiderata which evaluators should keep

in sight. If one or more maximum standards are not applicable to an evaluation, or could not be met, this should be disclosed and justified ⁽⁷²⁾.

Table 9: Preferred type of standards (minimum vs. maximum)

Preferred type of standards	Frequency	Percentage
maximum standards strongly prefer	5	29.4
maximum standards prefer	5	29.4
cannot decide	3	17.6
minimum standards prefer	2	11.8
minimum standards strongly prefer	2	11.8
Total	17	100.0

The majority of respondents preferred maximum standards. Only four endorsed minimum standards, two of them strongly, two less so. Three were undecided ⁽⁷³⁾.

The open question 12 asked what fundamental values evaluation standards should embody.

The most frequently mentioned values were participation with, cooperation between, and inclusion of, all stakeholders.

Transparency, integrity and frankness of the implementers were also often mentioned. Respondents also said that standards should provide leeway for adopting many different methods.

Some respondents regard reproducibility and transferability of findings as the central determinants of evaluation quality. The evaluation's findings should be useful and its impact beneficial.

The following criteria were mentioned: propriety; validity of findings; adoption of a long-term perspective with follow-up studies; stakeholder acknowledgement; implementation of a formative evaluation or a process evaluation; and competence of implementers and their responsibility for promoting the public good. Slightly over a third of the respondents did not answer this question. One participant feels that

most basic values are already incorporated in the JC standards.

Question 10 asked if current standards had significant gaps or omissions. Of the 19 respondents, 8 did not name any or did not answer the question.

Those who did answer usually felt the absence of a stipulation that the focus of studies should be generally extended, e.g. that indicators of success other than 'rate of employment' should be included in evaluations, or that studies should also feature variables like the macroeconomic and societal effects of evaluands or their social and educational environment.

Mirroring fears expressed frequently in the answers to question 9, some respondents desired additional emphasis on flexible and pluralistic evaluation approaches and consideration of cultural and historical idiosyncrasies. They also reiterated the principle that in each individual case evaluators should be free to make decisions according to their own ethical convictions.

Some respondents who play a major role in VET evaluations want them to be clearly directed towards supporting objectives and processes of vocational training, for example by more fully involving active participants:

⁽⁷²⁾ For details see the excursus on the meaning of the word 'standards' in Section 2.2.

⁽⁷³⁾ We could not detect a clear pattern for this answer. It did not correlate with the primary position in evaluations, proximity to the VET field or professional background.

- (a) '[...] content items, pedagogical items, items linked to management, teaching and support staff, items linked to participants (pupils, students, apprentices), physical resources, organisation [...]';
- (b) 'involvement of trainers, trainees, employers and other stakeholders in the evaluation process [...]';
- (c) 'recognition of the need to develop effective processes to evaluate intangible outcomes – e.g. the impact on individuals and communities in terms of quality of life, personal development, etc.; which make a real difference [...]'

This is an allusion to JC standard P1 on service orientation, which does not exist in the more general SEVAL and DeGEval standards, since it refers explicitly to the evaluation of education and training programmes: 'Evaluations should be designed to assist organisations to address and effectively serve the needs of the full range of targeted participants'. The explanatory notes on the standards contain the additional comment that the evaluations should help ensure that education and training objectives are appropriate, that learners' development is sufficiently heeded and that programmes which are useless or even harmful are abandoned. In this way evaluations can contribute towards making projects accountable to society and the community. Planners, implementers, users and participants must look beyond the interests of educators and organisations and aim to further learners' development and improve society as a whole. Evaluations should serve the interests of community programme participants and society. The JC guidelines on the standards explicitly state:

- (a) 'Evaluations should be planned which foster the quality of programmes for education, initial and continuing training.';
- (b) 'Evaluations should be used to identify intended and unintended effects of the programme on the learners.';
- (c) 'Teaching and learning processes should be interrupted as little as possible, but at the same time effort should be made to realise the evaluation project'.

We feel it would be useful to draft a correspondingly formulated VET standard ⁽⁷⁴⁾.

Other wishes were expressed by a few individuals: paying more attention to external consistency than internal; expounding the qualifications and experience of evaluators; incorporating long-term perspectives; establishing uniform basic terms and definitions, for the sake of international comparisons; and defining various sets of standards for the different capabilities of the implementers.

Over half the participants did not respond to question 11, which solicited alternatives to the standards that would improve VET evaluation quality. The rare suggestions that were made would primarily complement the standards rather than replace them. Examples include improving exchange between all stakeholders through regular conferences or establishing electronic communication networks. Also mentioned were introducing a system for certifying evaluators and/or institutes to guarantee their competence and developing certain aids (such as publishing survey guidelines) where possible.

The only possible alternatives to the use of standards that were proposed were social cost-benefit analysis and the capabilities and functionings theory (1987) devised by the 1998 Economics Nobel Prize winner Amartya Sen.

5.4. Summary of survey findings

This was the first survey on evaluation standards involving experts from most EU Member States. Despite the small sample and short questionnaire, the poll enabled us to identify tendencies and provided numerous stimuli for discussion on the further development of evaluation standards.

The sample mainly consisted of evaluators and researchers. The respondents were scholars in social and political science, the liberal arts and economics. Engineers and natural scientists were rare. The respondents identified themselves with a total of 13 different national professional cultures, giving the study a broad spectrum. The best-represented area was northern and western Europe. Around half the experts are familiar with VET but have had very little involvement in the

⁽⁷⁴⁾ See Summary and Outlook.

field. The remaining respondents described VET as a major or their main working field.

Everyone has a generally positive attitude towards evaluation standards. None of the respondents felt that standards do not matter or are unnecessary or even harmful. The best-known set of standards is the US joint committee standards for evaluation and the Guidelines for evaluators. The vast majority of the respondents named at least one set of standards with which they are at least familiar.

Respondents see the main benefits of evaluation standards as improvement in the quality of evaluations and an opportunity to make evaluators' work more legitimate and transparent. However, they fear that utilisation of the standards could restrict the plurality and flexibility of evaluations in theory and in practice, or that standards could be applied too rigidly.

When given the choice, the majority preferred maximum standards, which provide orientation, stimulate competent dialogue on evaluations and their methods and are open to innovation and further development. Only a few favoured precisely formulated minimum standards.

Respondents named involvement of all stakeholders and transparency and use of a wide

variety of methods as the most important hallmarks of evaluation standards. Correspondingly, well-known standard sets were felt to lack a standard which emphasises the desired flexibility and plurality of evaluation approaches and models. Most respondents did not see a superior alternative to evaluation standards and only suggested enhancements which concern evaluation management.

In summary, we can see that many respondents who agreed to take part in the survey due to an interest in the topic had already had some experience with evaluation standards ⁽⁷⁵⁾. The selection procedure (to some degree self-selection) could account for the very positive overall assessment of the standards (question 5). Objections to VET standards are equally applicable to minimum standards and are thus consistent with the fact that the majority of the respondents favoured maximum standards. Evaluation plurality seems to be an important fundamental value in European evaluation. On the one hand, this is explicitly stated in certain standards. On the other hand, erratic developments such as the inappropriately rigid application of evaluation standards could jeopardise it.

⁽⁷⁵⁾ See answers to question 6.

6. Reflections on VET evaluation standards literature

Documentation research involved a search for pertinent articles from the last five years on the quality of VET evaluations and evaluation requirements. Reflection can take place at the end of an evaluation or from a scientific/methodological perspective during evaluation research. Older literature has only been consulted when it is perceived as being particularly relevant or regarded as a standard work in this field.

Although evaluation methodology has most of its roots in North America, where it is widely used and has a long research tradition, the literature to be assessed should stem from European authors or reflect a European background. This should ensure that European cultures and unique institutions receive appropriate attention. This approach should prevent the unqualified import of American evaluation culture, which could reduce acceptance and provoke resistance. Mistrust of government intervention and a public right to information characterises American evaluation culture (Schmidt, 2000). Empirical social science research is common in the US. In Europe, however, labour-market policy studies emphasise different programme outcomes by comparing employment and income effects. European countries rarely conduct empirical social science research. We should observe separate innovations in European States (Toulemonde, 2000).

Literature in English and German is systematically researched. French and Italian sources are consulted in exceptional cases.

The process involved methodical evaluation of reference databases, particularly those of Cedefop, the University of Osnabrück Library (comprehensive social science section) and the University of Cologne Library (designated as German Economic Science Library) and Internet research on evaluation standard terms.

Perusal of the literature and subsequent categorisation of text segments according to individual standards consistently show that the standards overlap. Notes on evaluation quality requirements, therefore, cannot always be clearly assigned to a single standard. Below, evidence of

use is cited under the most applicable standard; reference to overlap is made where necessary. Overlap also stems from the fact that some standards are either directly or indirectly related. The individual standards belong to very different analytical levels and, consequently, the number of comments on each standard differs markedly.

The literature analysis reveals that development of the theoretical basis for evaluation has slowed during the past decade ⁽⁷⁶⁾. It has given way to a phase of consolidation and application of established evaluation models. One field of application is VET. Literature on VET evaluation encompasses a broad palette of perspectives. Some articles focus on model theory or evaluation methods, while others report on completed evaluations. Another clear trend is documentation which provides guidance on conducting VET evaluations. We can define various levels within the VET evaluation literature examined, covering evaluations on supporting government decisions or improving the quality of individual programmes and evaluations as a mechanism to initiate public VET debate. Evaluands in VET literature range from the use of new media in continuing training, through vocational training pilot projects, to individual programmes as part of in-company continuing training. Assessment of evaluation standards should incorporate the whole spectrum of possible VET evaluands.

When dividing VET into micro-, meso- and macro-perspectives, we can generally assign these levels to different reference disciplines. Economics tend to dominate the macro-perspective. Economists often use quasi-experimental investigation forms or 'advanced' quantitative procedures. Economic theories such as the human capital theory are employed to try to explain many meso-level phenomena. However, sociological and educational methods and theories may also apply, depending on the line of investigation. The micro-level is primarily viewed from a psychological or educational perspective. Standards which are to apply specifically to VET should therefore comply with key scientific criteria in all reference disciplines.

⁽⁷⁶⁾ See Stufflebeam (2001) for a survey of diverse evaluation models over the past decades.

Various evaluation studies exist for the levels distinguished here. Universities, related institutions and individual researchers conduct macro-evaluations as a rule. They usually observe and comply with all general scientific (methodological) standards, some of which feature in the evaluation standards, as a matter of course. Academic expertise is much less evident at meso- and micro-levels. For example, part of the job of staff developers is to conduct evaluations. They have insufficient methodological training for this task. Clear standards could act as guidelines with an initial and continuing training function for this group in particular.

Not all standards are equally applicable to every evaluation project, which is also true for VET evaluations. Nevertheless, the validity of each individual DeGEval standard has been confirmed in various VET contexts. The literature studied contains concrete quality requirements, advice and guidance on using evaluations which resemble the individual DeGEval standards. We can illustrate the individual standards in terms of the VET evaluand and its characteristics and we can refine some of the standards further. We will therefore proceed by briefly introducing each group of standards and adding a commentary on individual standards. This commentary is partly illustrative and descriptive, and partly more reflective, depending on the conflict potential which each standard contains.

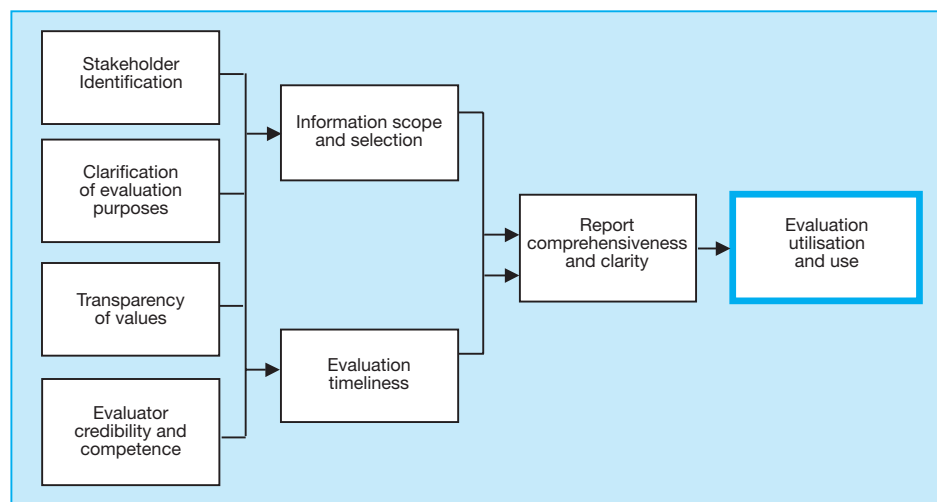
Below are the 25 DeGEval standards, grouped according to the four standard categories, with notes gleaned from European VET-oriented literature. We have refrained from providing a detailed description of each standard, as this appears in the attached version of the DeGEval standards and their explanatory notes (printed in the Annex).

6.1. Commentary on the utility standards

‘The Utility Standards are intended to ensure that an evaluation is guided by both the clarified purposes of the evaluation and the information needs of its intended users’ (DeGEval, 2002, p. 8). The utility standards are particularly relevant when interfacing with the (intended) users of evaluations and their findings. Stakeholders may be (programme) managers or employee representatives. Evaluator competence depends on experience in the field of enquiry and intercultural awareness. In view of the vast disparities, discussion of values is particularly important not only within various target groups, but also in different European countries. The utility standards seem to be relatively sensitive to cultural diversity.

Standards N1/U1, N2/U2 and N5/U5 help clarify the basis and hence the interests, influences, purposes and values of a specific evaluation.

Figure 5: **Seven points which make evaluations useful**



Source: author's representation

This is linked to evaluator competence and credibility (N3/U3). These standards must also be respected in the following operational planning steps. These include information scope and selection (N4/U4) and report timeliness and dissemination (N7/U7). They should be brought together into a comprehensive and clear report (N6/U6) and result in a high usage of evaluation (N8/U8).

6.1.1. N1/U1: stakeholder identification

‘Persons or groups involved in or affected by the evaluation should be identified, so that their interests can be clarified and taken into consideration when designing the evaluation.’

Evaluations generally involve a concerted effort on the part of all participants, especially in initial and continuing vocational training. They share responsibility for the evaluation process. As a rule, all parties discuss evaluation planning, implementation and findings. ‘Goals [in the context of evaluations] [...] are often negotiated between at least three interest groups, namely between management and works council representatives on the one hand and researchers on the other’ (Antoni, 1993, p. 315). This applies to evaluation of both in-company and extra-plant vocational training. Many different groups of decision-makers and other stakeholders exist, particularly in training partnerships. The EU also stipulates that ‘involvement of the social partners in all phases of the evaluation is crucial for finding viable ways of meeting the prevailing local labour-market requirements and solutions to employment issues’ (Gontzou, 1997, p. 62).

According to Reischmann (2003), evaluations must not muzzle people if adult education is to encourage them to become independent and active citizens, responsible employees and well-rounded personalities. Participants should have the opportunity to play an active role in the assessment process and to take advantage of this aspect of learning within evaluations to nurture their own development.

For evaluation of in-company training measures in large enterprises, the evaluator should develop tools in consultation with the head of the training department and/or the head

of the personnel department, in cooperation with other department heads (Tremea, 2002). The role of trainers in the evaluation should be clarified. They could participate in the evaluation by observing, measuring performance or identifying areas where more training is needed. The form of trainee involvement in the evaluation process must also be established, e.g. completion of questionnaires, interviews, self-assessment. Increases in productivity as a result of training measures could be determined by consulting external parties interacting with the company. These could include suppliers, distributors, current and potential end users, employers’ associations, trade unions, etc.

Stakeholders for evaluation of pilot projects in school-related education include the pupils and teachers of government schools, educational research institutes and their various departments (e.g. vocational school departments) and academic institutes providing educational support, as well as the pilot project sponsors, such as the central education ministry, a regional ministry or a national vocational training institute. ‘These people always have divergent interests. Their allegiances are to quite different institutions. This leads to contrasting interpretations of a pilot project’s mandate and widely varying concepts of their own function, their role in the pilot project and the functions and roles of other players’ (Sloane, 1995, p. 13).

Stakeholder orientation is also dictated by culture, particularly relating to hierarchy or egalitarianism. In essentially egalitarian societies, interpretation takes for granted the incorporation of different stakeholders in the evaluation process. In other societies a certain degree of unequal empowerment and strongly differentiated spheres of influence are both legitimate and desirable (Taut, 2000). Reflecting on the American standards, Jang (2000) notes that in South Korea it is common practice only to consider the expectations of the commissioner. Although the cultural diversity within the EU is certainly not as great as between the US and South Korea, differences between European countries also have to be taken into account ⁽⁷⁷⁾.

⁽⁷⁷⁾ Hofstede’s (1980) investigation of the ‘power distance’ dimension showed that it is relatively high in Belgium, Greece, Spain, France and Portugal in comparison with other European countries.

6.1.2. N2/U2: clarification of the purposes of the evaluation

'The purposes of the evaluation should be stated clearly, so that the stakeholders can provide relevant comments on these purposes, and so that the evaluation team knows exactly what it is expected to do.'

Antoni (1993) maintains that different interest groups and their varying goals sometimes obscure the evaluation purpose. Nevertheless, the purposes of an evaluation should be explained in accordance with standard N2/U2.

Possible evaluation purposes are preparation of government decisions, preparation of company decisions on training, information on individual decisions and quality improvement for specific programmes ⁽⁷⁸⁾.

'At its most effective, the evaluation process needs to relate to the needs and objectives of the organisation, its component parts (e. g. departments or teams) and the individual employee. It should be recognised that the requirements, and therefore the objectives, may be different for each of these. Provided this is recognised, and the expectations from the training or development activities are recognised, then the needs of all three can be accommodated. There should be a coherent structure in the evaluation process that starts with expectations, leads through to reaction and measures the changes' (Field, 1999, p. 218). The evaluation purpose should, therefore, coincide with the goals of organisation units or the corporate strategy.

A survey of 2000 enterprises in Europe on the purpose of training evaluations revealed the following points (Field, 1998b, p. 72; authors' additions in parentheses):

- (a) to measure the extent to which objectives have been met (Q);
- (b) to encourage the effective use of resources (P);
- (c) to further develop individuals and their careers (P);
- (d) to improve the organisation's turnover (P);
- (e) to increase the organisation's competitiveness (P);
- (f) to obtain feedback on the training provision (Q);

- (g) to identify the impact of the training activity on the employee's job performance (Q);
- (h) to justify money spent on training (P);
- (i) to identify the contribution to business objectives (Q);
- (j) to identify the contribution to organisational performance (Q);
- (k) to measure the effectiveness of the training (Q);
- (l) to provide information for sponsors (P).

This list demonstrates the difficulty in distinguishing between evaluation purposes (P) and questions (Q) which the evaluation should answer ⁽⁷⁹⁾. Purposes describe something which an evaluation should set in motion in the social and economic environment. Questions describe something which the evaluation should clarify (N4/U4).

Ideally, all evaluation processes should disclose and explain the evaluation purpose to the stakeholders. Everyone should know what will happen to the survey data and what kind of feedback they can expect (Field, 1998b, p. 25).

6.1.3. N3/U3: evaluator credibility and competence

'The persons conducting an evaluation should be trustworthy as well as methodologically and professionally competent, so that the evaluation findings achieve maximum credibility and acceptance.'

The competence of evaluators is a significant factor, since they do not have a standardised job profile. Independent evaluation and evaluation research courses in social sciences are relatively rare at European universities. A few European countries offer postgraduate courses ⁽⁸⁰⁾.

Application and interpretation of existing tools requires empirical and methodological knowledge. This applies all the more to tailoring of tools. In addition to these methodological skills, also cited in the accuracy standards, evaluators must demonstrate knowledge of the evaluand and its context. 'Professional competence as an evaluator in the technical field and geographical area of the project is one of the principal elements in the selection criteria for designating

⁽⁷⁸⁾ For more details, see Grubb and Ryan (1999), pp. 21 f.

⁽⁷⁹⁾ Confusion about evaluation purposes and programme goals is not uncommon.

⁽⁸⁰⁾ E.g. Spain, Sweden and Switzerland.

the evaluation team members. Objectivity and independence are the other key considerations for selecting evaluators. The degree of independence, however, depends on who designates the evaluators' (ILO, 1999, p. 10). Evaluators who work exclusively in VET and show outstanding expertise in this field are in danger of becoming blind to programme malfunctions and positive side-effects.

It would be sensible for an (internationally active) evaluation team to include at least one evaluation specialist and one VET expert, and other people with knowledge and awareness of the economic and social needs and problems of each country in which the evaluation takes place (Grubb and Ryan, 1999).

Other authors, such as Wottawa, emphasise that academic competence must be transferred to corporate practice (Wottawa, 1999, pp. 112-113). 'The "academic background" (education, social sciences, psychology, economics) of potential evaluators is secondary. The business world is less interested in which disciplines employees have training in, focusing more on whether they show practical competence which transcends subject boundaries. For most vocational training evaluation procedures it is vital to work with people from diverse specialist backgrounds. Many evaluation projects must integrate academics, education experts, management and participants themselves. They all have different educational backgrounds.'

Evaluators must also be aware of the limits of their own knowledge and skill. This could prompt them to consult other experts and delegate certain duties to other parties. VET could, for example, involve determining the motivation of participants in a measure. If an evaluator's psychology expertise does not suffice for this, it makes sense to apply standardised procedures or to leave the collection and interpretation of findings to other psychologists (Tremea, 2002). The same applies to determining psychological profiles, which are also very sensitive.

In deciding who should conduct the evaluation, one must also consider whether it should be internal or external. Schmidt (2001) argues that external evaluation is advisable, since programme planning could be based on false

premises. The competence of an external evaluator could be helpful, and outsiders are more scientific and independent.

If evaluators conduct evaluations in unfamiliar countries, cultural distance will play a role. However, they will also have interpersonal distance from the people in the country concerned. This can be advantageous, enabling them to assume a 'balanced view' (Hendricks and Conner, 1995). If evaluators work abroad, intercultural skills may be part of their qualifications. Sensitivity to social, cultural and economic differences between the various stakeholders is crucial.

How different cultures determine evaluator credibility can vary dramatically. A society with a valid 'seniority principle', for example, may automatically regard the older generation as the more, or even only, competent group (Jang, 2000). Social status and gender can also significantly affect assessment of evaluator competence and credibility, depending on the culture.

6.1.4. N4/U4: information scope and selection

'The scope and selection of the collected information should make it possible to answer relevant questions about the evaluand and, at the same time, consider the information needs of the client and other stakeholders.'

The logic model is one tool which can be applied to clarify objectives and structure the programme for evaluation ⁽⁸¹⁾. This specifies overall goals, interim goals, indicators and effects and puts them into context. The tool is widely used in evaluations for structuring internal programme logic and formulating questions to be addressed by the evaluation.

An oft-cited and popular approach for detailing VET evaluation questions is Kirkpatrick's (1994) four-level model. This first examines learner reactions, and then what participants have gained from the programme. The third stage evaluates behaviour in the work environment, and the fourth studies the results from an organisational perspective. This last stage entails a return on investment. Thus we have various evaluands. It would no doubt be more sensible to establish the evaluation purposes (N2/U2) before formulating questions or indicators.

⁽⁸¹⁾ The 'logic model' is often used to structure evaluations.

‘This obliges evaluation providers to consider the information needs of decision-makers (in business, not in research) even more closely when selecting their strategies and assessment indicators. If this does not happen, there is a risk that decision-makers, who ultimately provide the funding, will opt for alternatives, i.e. at best other evaluators or, at worst, even to dispense with scientifically sound evaluations altogether’ (Wottawa, 1999, p. 108). The information purpose determines the value of knowledge, and not the quantity of information, according to Weiß (1997, p. 108). Information for evaluations should be chosen and condensed in such a way that it can serve as a basis for decision-making ⁽⁸²⁾.

6.1.5. N5/U5: transparency of values

‘The perspectives, procedures and thought processes that serve as a basis for the evaluation and the interpretation of the evaluation findings should be described carefully to clarify their underlying values.’

‘Before undertaking the mission, the team members should also familiarise themselves with the cultural and social values and characteristics of the recipients and intended beneficiaries.’ The ILO Guidelines concur (ILO, 1999, p. 12). Cultural values can vary dramatically within a country and between companies and organisations. This standard on identification of values is highly relevant to evaluations which encompass several European states or which are conducted in different European countries. Standard N3/U3 also applies, as intercultural competence strongly influences the identification of values.

‘Naturally, integrative concepts will spark considerable debate as to their explicit value judgements with regard to the weighting of different types of outcomes as well as the time preferences or even group preferences.’ (Schmidt, 2001; p. 9). Trade-offs can occur between different times or different social groups. Selection of individual parameters for evaluations is vulnerable to biases, as it can affect the significance and even determine the survival or demise of political and corporate programmes.

6.1.6. N6/U6 – report comprehensiveness and clarity: evaluation reports should provide all relevant information and be easily comprehensible

Annex 2 to the guide to evaluation of EU expenditure programmes (European Commission, 1997) formulates specific questions for assessing the quality of evaluation reports: ‘Is the report well presented? [...] Is the scope of the report adequate? [...] Is the methodology of the report appropriate? [...] Are the report’s conclusions and recommendations credible?’ (see also G8/A8)

These key questions are elucidated further. For example, the last question is complemented by the following additional questions. ‘Are findings based firmly on evidence? Are conclusions systematically supported by findings? Are recommendations adequately derived from conclusions?’ They not only articulate requirements for the form of the report and the style of presentation, but also impose clear quality demands on the content. This overlaps with considerations of methodology and data quality in other standards ⁽⁸³⁾.

In the context of vocational training pilot project research, Zimmer (1998, p. 598) comments that the findings should be processed in such a way that other enterprises and training institutions can benefit from them. It should be possible, therefore, to transfer available interim results, and not just conclusive findings, to other companies or training establishments with similar problems. According to Kaiser (1998), the potential for gaining scientific insights from pilot projects depends in particular on the structure and presentation of texts produced in the course of the scheme, such as the final report and project documentation on teaching and learning arrangements. Addressees, teachers, school administrators, trainers, company managers, education policy-makers, cultural bureaucrats and education coordinators will undoubtedly read a final report on a pilot project only if it is not too extensive and overloaded with details and jargon. Every pilot project team should consider each time how to compose the final report to ensure

⁽⁸²⁾ It goes without saying that there are other evaluation purposes besides providing a basis for decision-making, such as ongoing improvement or accumulation of general knowledge.

⁽⁸³⁾ See the accuracy standards G1/A1, G2/A2, G3/A3, G4/A4 and G8/A8.

that important findings and results are accessible to vocational training policy-makers and useful to vocational training research. Reports should be tailored to the relevant target group (see also user groups N1/U1).

An evaluation report should also contain a list of any problems regarding concepts, contents and methods which may surface (Kaiser, 1998, p. 547). This relates to the accuracy standards and is crucial to the subsequent meta-evaluation.

6.1.7. N7/U7: evaluation timeliness

‘The evaluation should be initiated and completed in a timely fashion, so that its findings can inform pending decision and improvement processes.’

Ideally, the report should be completed immediately after gathering data. Often deadlines are based on the needs of third parties, such as data for important meetings in which results are presented (Field, 1998b, p. 12).

Moreover, it has been ascertained in connection with N7/U7 that evaluation design and quality heavily depend on the timing of evaluation planning. Evaluations planned after the launch of a programme lack certain opportunities to influence the evaluation design to ensure evaluability and allocate participants to test and control groups. This applies especially to experimental studies. But before-and-after comparisons cannot be accurate if evaluation planning only commences after the start of a programme. Beginning evaluation design only after the decision to run a programme, after the successful launch of the programme or even after its conclusion are common occurrences in VET ⁽⁸⁴⁾. In such cases it is possible to complete the evaluation report in good time, but the evaluation itself cannot begin punctually. This affects both evaluation content and method.

Here we must note that ‘timeliness’ of the evaluation as described in the text to the DeGEval standards can favour the production of quick results. Most evaluations at the end of a VET programme study short-term effects appearing in 30 to 90 days. Evaluations which measure effects after two years tend to have more complex, often

randomised designs. However, since short-term and long-term effects are not necessarily related, a longer perspective is needed. Evaluations which only observe short-term developments may approve programmes with immediate impact and underestimate those whose effects only become evident or mature after several years. Focusing on immediate benefits can hamper observation of long-term effects. The potential worth of a programme for vocational training can increase or decrease over the course of time.

Gaude (1997, p. 55) hypothesises that the income of former vocational further training measure participants could be higher after a period of job-seeking than that of the control group. The increased competence of the former trainees would permit them to reach a higher rung on the career ladder. However, they may not have upgraded their qualifications and could also stagnate in poorly paid jobs. These possible effects can only be observed and measured over several, longer survey periods. Gaude states that many evaluations do not last long enough to gather this data. Extending the evaluation over several years is the only solution. Grubb and Ryan (1999) propose five to six years.

Fay (1997, p. 111) also calls for longer evaluation periods, especially for training programmes. The following relationships could be of interest (Tremea, 2002): training and employment, training and promotion, training and job-keeping. These questions could be helpful. Are the former participants employed in the occupation for which they trained? Do training participants use training lessons regularly? What training would participants have needed to perform their current duties more efficiently?

Calls for longer evaluation periods increase evaluation complexity and costs. Furthermore, it takes longer to publish final evaluation reports. Stretching evaluation periods can also encourage the separation of programme evaluation from political cycles. Programme and evaluation duration should be interdependent. The duration of a programme’s potential impact, including multiplier effects, also has a close bearing on evaluation duration and timing of surveys.

⁽⁸⁴⁾ Expert discussion in the Vocational and in-company continuing training task force at the DeGEval conference in Mainz on 17 October 2002.

6.1.8. N8/U8: evaluation utilisation and use

'The evaluation should be planned, conducted, and reported in ways that encourage attentive follow-through by stakeholders and utilisation of the evaluation findings.'

Meta-evaluations are conducted to establish how VET evaluations are utilised. We know of no European studies on this subject.

A survey of enterprises in Europe obtained the following responses to the question of training evaluation use (Field, 1998b, p. 73):

- (a) facilitating and reflecting on the transfer of learning to the workplace;
- (b) reducing staff turnover;
- (c) ensuring that training meets company and individual objectives;
- (d) raising awareness of the benefits of training;
- (e) increasing staff motivation;
- (f) improving the effectiveness of training activities;
- (g) measuring productivity increase;
- (h) increasing individuals' responsibility for their own training and personal development;
- (i) involving managers in the training and evaluation process.

This list demonstrates that both the evaluation process (goals become clearer, broken links in the chain of training elements are discovered and repaired, etc.) and its findings (well-founded decisions are made, which increases both worker motivation and productivity in the long term) can trigger the stated forms of utilisation. Some evaluation approaches prefer process use (e.g. the qualitative approach related to organisational development) while others adopt the findings use (as in quasi-experimental approaches). As standard N2/U2 shows, stakeholder use requirements dictate prioritisation of the central evaluation benefit. Models and methods should adapt to use requirements, not vice versa. More academic evaluators and pragmatic evaluation approaches often clash.

Evaluations of VET measures should provide commissioners with clear instructions and comments not only to demonstrate that a training programme was completed more or less successfully, but also to encourage further utilisation of the findings. This applies particularly to formative evaluations. They should formulate specific proposals for possible changes, as 'the

purpose of evaluation of training is not to prove, but to improve' (Tremea, 2002). Evaluation stakeholders such as training measure purchasers, training measure providers, training participants, must be activated. Potential evaluation utilisation should perhaps encompass a wider group to ensure that colleagues of participants, or entire departments, enterprises or organisations, are also informed. The actual or unrealised benefit of an evaluation and the (non-)application of proposals should be recorded (follow-up). Was the training programme restructured on the basis of the evaluation? Did selection of trainers reflect the previous evaluation? If additional training was recommended, has it already taken place?

Evaluations can strengthen interpersonal relations and worker motivation within a company. Workers who may also be training participants could appreciate colleagues listening to their opinions and adopting their ideas, if there is keen interest in the results of a training initiative (Tremea, 2002). In the longer term it is vital to utilise the information gathered conspicuously to motivate workers to participate in future evaluations.

Several factors can bolster the use, and hence the success, of evaluations within enterprises and organisations. A company with a corporate culture based on trust rather than mistrust promotes training as an investment. This kind of environment also tends to support data compilation and utilisation. Linking evaluations to relevant strategic and organisational goals increases the probability that recommendations will be respected and implemented (Field, 1998b, p. 75). The attitude of senior management to, and support of, training and its evaluation is a deciding factor in the utilisation of evaluations and their findings.

Efforts to establish continuity can also boost evaluation utility. This includes concurrent development of monitoring systems ⁽⁸⁵⁾ which can supply data for evaluations and channel the information obtained in evaluations into the monitoring process. This applies particularly to state-financed and state-run initial and continuing training activities.

It should be emphasised that evaluation activities are not always beneficial. They can be worthless or even harmful. Reischmann (2003) coins the merit criterion of 'didactic utility' specifically

⁽⁸⁵⁾ See Auer and Kruppe (1996) for an overview of monitoring systems in the EU.

for adult education. He maintains that evaluations can only help improve the understanding and structure of adult education if they apply this criterion from the outset. Reischmann attaches more weight to this factor than to any other. He states that evaluations are only a valid aspect of adult education if their andragogical and didactic intentions and consequences are clear.

6.2. Commentary on the feasibility standards

‘The Feasibility Standards are intended to ensure that an evaluation is planned and conducted in a realistic, thoughtful, diplomatic and cost-effective manner’ (DeGEval, 2002, p. 9).

The feasibility standards are highly relevant to formative evaluations. Diplomatic conduct is especially adaptable to different cultures and thus more sensitive to national evaluation environments than the other two feasibility standards. Both appropriateness of the procedure employed (D1/F1) and diplomatic conduct (D2/F2) significantly affect evaluation efficiency (D3/F3).

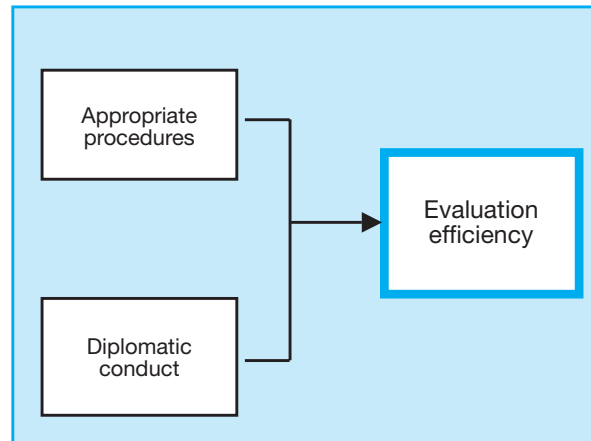
6.2.1. D1/F1: appropriate procedures

‘Evaluation procedures, including information collection procedures, should be chosen so that the burden placed on the evaluand or the stakeholders is appropriate in comparison to the expected benefits of the evaluation.’

Decisions on the evaluation design must reflect the type of programme being assessed and the nature of the programme’s expected impact, which is affected by the following programme attributes (Lindley, 1996, pp. 853-854):

- (a) scale:
 - (i) the coverage of the programme relative to the size of the socioeconomic space for which the evaluation is being conducted;
 - (ii) the extent of the tax expenditure involved relative to the costs perceived by the actors whose behaviour is being influenced;
- (b) selective: dealing only with a broad section of economic activity, whether distinguished by aggregate sector (e.g. agriculture or manufacturing), spatial area (e.g. poorer nations or regions) or major socioeconomic group (e.g. women);

Figure 6: **Two points which make evaluations feasible**



Source: author’s representation

- (c) targeted: focused more sharply on particular sectors (e.g. coal mining), subregions (e.g. level 2 of the Eurostat regional classification) or labour force groups (e.g. unemployed young people, women returning to the labour force);
- (d) transitory: where the policy is seen by the actors as being merely a temporary measure, or one which may be used only recurrently from time to time;
- (e) countercyclical: where policy intervention is a reasonably predictable form of countercyclical measure (rather than being considered to be so *ex post*);
- (f) long-term: where the policy intervention is seen to be a long-term measure, even though aspects of it may be subject to variation according to socioeconomic conditions.

At one end of the scale, evaluands can be relatively small, target a specific group and cover a limited period. The other extreme comprises extensive, long-lasting programmes with diverse, sometimes hierarchically-related target groups (e.g. provider managers, trainer trainers, trainers, end consumers such as young people and their parents). Both extremes – and all graduations in between – require different evaluation models and methods.

Quantitative, standardised tools, which necessitate considerable investment in development or adaptation, may be efficient for large programmes. Qualitative but flexible, practical tools are often more appropriate for small programmes.

A common situation involves nationally or even European-funded programmes which are imple-

mented in many locations and function almost independently. This raises the question of whether the combination of blanket monitoring and local case studies, quasi-experiments using control or comparison groups, or a cluster evaluation is the most suitable evaluation design (Beywl et al., 2003).

Butz (2000, p. 432) maintains that assumptions on the supposed acceptance by those questioned and pollsters should steer the selection or construction of the actual survey materials. If programme organisers, for example, are involved in a dense, binding monitoring system, they will resist additional, written surveys, but are more likely to accept telephone interviews or interactive group survey procedures with integrated exchange. Reischmann (2003) also advises omitting everything which will not be evaluated if information is obtained directly from participants. It is important to ascertain whether data compilation can be spread among various (groups of) people so that nobody is overtaxed. It is also prudent to check whether material or documents already contain some necessary information which does not have to be gathered separately.

6.2.2. D2/F2: diplomatic conduct

‘The evaluation should be planned and conducted so that it achieves maximal acceptance by the different stakeholders with regard to evaluation process and findings.’

Views on appropriate diplomatic conduct depend on national cultures and differ between commercial and non-profit organisational environments. Even the term ‘diplomatic’ has divergent or even contradictory connotations (cautious, adept, covert, indirect, manipulative, etc.), depending on the culture ⁽⁸⁶⁾. This alone indicates the standard’s high cultural sensitivity. At the same time the standards N1/U1 and F2/P2 are relevant.

Practice shows that resistance from employee associations such as works councils can thwart evaluation projects. Data protection officers can also exert a strong influence in Germany. To avoid unexpected barriers, employee representatives in an enterprise or organisation, who are granted a voice under national law, should be involved as

extensively as possible in planning from an early stage.

Resistance may stem from negative experiences of preservation of anonymity and confidentiality in previous surveys. For example, unofficial but widely distributed evaluation documents may contain the names of individual trainers. Or, particularly in small organisations, it is possible to deduce who assessed performance negatively or positively, as presentation of data in the final report is too specific.

At worst, this can result in a warning, transfer or termination of contract for the affected programme organiser or trainer, despite assurances of anonymity and confidentiality. Such occurrences deter enterprises and organisations from participating in evaluations. To avoid this effectively, Butz (2000, p. 437) recommends involving works councils, data protection officers and staff representatives from the evaluation design stage. It may be wise to conclude a written evaluation agreement with the works council. Names of participants and other individuals should not be mentioned in public evaluation documents. Reports on smaller departments should be summarised.

6.2.3. D3/F3: evaluation efficiency

‘The relationship between cost and benefit of the evaluation should be appropriate.’

Nuissl (1999, p. 73) notes, ‘It is necessary to develop an acceptable system of evaluation, assessment and monitoring to assess the overall effectiveness of projects and to ensure the quality of the outcomes.’ When commissioning, tendering for, planning and implementing an evaluation, one must ensure that the invested resources are economically proportional to the expected use of the evaluation. This concerns personnel involvement in the evaluation as well as the costs and burdens which the enterprise or other organisation hosting the evaluation may incur through data collection or supervisory panel meetings.

Decision-making on the overall investment an evaluation warrants should consider the planned scope of the evaluation findings (Tenberg, 1998, p. 533). For a pilot project it may be sensible to overscale the evaluation to the entire programme

⁽⁸⁶⁾ The titles of JC standard F2, Political viability, and SEVAL standard D2 of the same name often provoke ambivalent reactions, as political is associated with unfounded, irrational or arbitrary in enterprises.

costs, as subsequent transfer of the project will affect a large number of participants and demand a correspondingly high budget. Company-level evaluation of every kind of continuing training is superfluous. Only one or two employees may participate in a training programme, or minimal investment has been needed, or all parties without exception are convinced of the use of a well-established training programme. In these and similar cases an evaluation will not be conducted or will only take place on certain levels.

The PAVE project's evaluation resource pack asks companies the following questions to help them decide whether an evaluation should be conducted, and if so, on what scale (Field, 1998b, p. 52).

- (a) how many people does the training or development measure affect?
- (b) how crucial is achieving the expected training goal for the company?
- (c) how likely is this measure to run again?
- (d) has the training provider received a contract before?
- (e) is the type of training new to the company, e.g. new communication technique, new skills?
- (f) to what extent do the training measure and the evaluation process support other areas of corporate policy?
- (g) is evaluation of the training or development measure (urgently) needed?

Weiß (1997, p. 107) states that the more precise the tools and the more differentiated the measurement criteria, the higher the investment. Commissioners and evaluators should consider what level of perfection is required. In practice, it will often be necessary to compromise between the desire for accuracy and the resources available. This applies both to individual programmes and to selection of sub-projects for evaluation (Lindley, 1996).

6.3. Commentary on the propriety standards

'The propriety standards are intended to ensure that in the course of the evaluation all stake-

holders are treated with respect and fairness' (DeGEval, 2002, p. 9).

Model contracts and legal foundations provide clear points of reference for the first two propriety standards, formal agreement (F1/P1) and protection of individual rights (F2/P2). Complete and fair investigation (F3/P3) and unbiased conduct and reporting (F4/P4), in contrast, are much harder to clarify and judge in the fierce conflict of interests. The form of the published findings (F5/P5) should embody the result of the precautions taken in the first four standards. Are all pertinent findings published or only those which do not collide head-on with the interests of key participants? This crucial decision should be made as soon as possible in the course of an evaluation, formally agreed (F1/P1) and communicated to the stakeholders (D2/F2).

The propriety standards set out the industrial relations requirements for VET evaluations, such as decision-making regulations and data protection. They also articulate cultural differences in treatment and protection of minorities ⁽⁸⁷⁾.

The propriety standards impose specific demands on evaluators' legal knowledge and social awareness (N3/U3). This is particularly important when evaluators work abroad.

However, the service orientation standard in the American JC standards (JC-P1) could be relevant to VET evaluations. The DeGEval standards do not contain this standard, as they are designed to apply beyond the field of human services (Section 6.5).

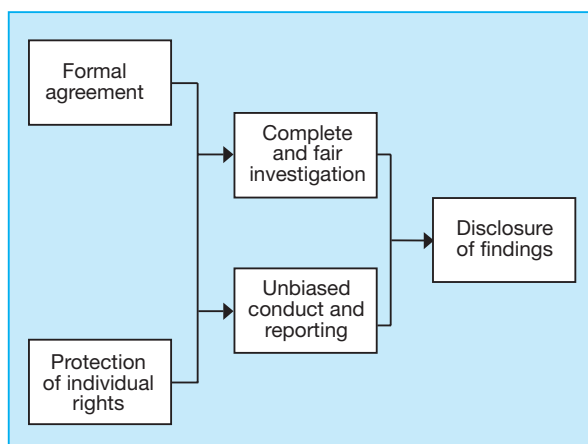
6.3.1. F1/P1: formal agreement

'Obligations of the formal parties to an evaluation (what is to be done, how, by whom, when) should be agreed to in writing, so that these parties are obligated to adhere to all conditions of the agreement or to renegotiate it.'

Evaluating EU Expenditure Programmes (European Commission, 1997) and the MEANS handbook (European Commission, 1999b, Vol. 1, p. 76) list key elements which a contract should normally contain: 'the legal base and motivation for the evaluation, the future uses and users of the evaluation, a description of the programme to be evaluated, the scope of the evaluation, the main evaluation questions, the methodologies to

⁽⁸⁷⁾ We are unaware of any study which appraises the various existing stipulations regulating fair and legal implementation of VET evaluations in EU Member States and general data compilation in the various subsystems (enterprises, public authorities, schools ...).

Figure 7: **Five guidelines which keep evaluations on a straight course**



Source: author's representation

be followed in data collection and analysis, the work plan, organisational structure and budget, the selection criteria for external evaluators, the expected structure of the final evaluation report' (European Commission, 1997, p. 38 f.). The JC standards include detailed guidelines on this (JC, 1994, p. 88).

We know of no lawsuits between commissioners and contract recipients in Europe to date which have appealed to the standards. However, this could be a future role of the standards, as has always been intended. When in doubt, courts will consult professional standards. The formal agreement should explicitly state whether the evaluation standards form the basis of evaluation implementation to create clarity between commissioners and contract recipients. We found no specific references to formal agreements on evaluations in VET literature.

6.3.2. F2/P2: protection of individual rights

'The evaluation should be designed and conducted in a way that protects the welfare, dignity and rights of all stakeholders.'

Initial or continuing training participants often have a high stake in their programmes. They are counting on obtaining a vocational qualification, which will open the door to certain professions, a livelihood and social status. When (re-)entering the world of work, continuing training participants can achieve promotion and secure their jobs, but they may also lose out if they are transferred, or their contract is not renewed or is terminated. Full-time VET staff and freelance training

providers, in particular, associate evaluations with great opportunities and high risks.

Personal data protection and appropriate handling of performance data and findings which can be traced back to individuals should be a priority in VET evaluations. For example, an evaluation may require psychological profiles, such as measurement of intelligence or other personal traits, with especially sensitive information. It is untypical for evaluations to gather this kind of data. If they do, to assess the aptness of a training concept to participants' initial cognitive status or to explain learning difficulties, for example, confidential handling of this data is vital (Tremea, 2002).

In any case it must be emphasised, ideally in the formal agreement (F1/P1), that neither the grading of participants nor the assessment of trainers is the aim of programme evaluations. There are independent standard sets for this. They prescribe much more precise and narrow regulations for protecting personal rights than the Programme Evaluation Standards (JC, 1988, Gullickson, 2002).

The gender issue is another important aspect. The status of men and women is culturally dependent and varies throughout Europe. Evaluations do not presume that a certain gender leads to better or worse training results. However, some occupational groups tend to employ mainly men or mainly women. Moreover, different European countries have diverging views on the role of women in the workplace. An evaluation must consider these aspects and decide whether or not to record participant gender.

The same applies to data on participant age. Older employees have more difficulty finding employment in some European countries than in others. In Scandinavia, for example, age tends to have less effect on the probability of finding a job. Ethnic or minority composition of a training group can also affect participant chances of employment. Evaluators should only gather or assess such sociologically and politically sensitive data if commissioners expressly request it and explain why (Tremea, 2002).

This highlights the cross-reference to the transparency of values standard (N5/U5), which should be respected at the conception stage of data collection in VET evaluations spanning national boundaries.

6.3.3. F3/P3: complete and fair investigation

‘The evaluation should undertake a complete and fair examination and description of strengths and weaknesses of the evaluand, so that strengths can be built upon and problem areas addressed.’

Identifying and eliminating weaknesses during evaluation implementation is conceivable. In practice this is more helpful than waiting to make changes until publication of the final report. Reischmann (2003, p. 253) maintains that in extreme cases, a final report could include the following: ‘We have identified the following weaknesses: [...] We employed the following measures to eliminate them successfully and permanently: [...] The evaluation report thus has no further recommendations!’ However, changes which have already been implemented should still be identified and documented in detail.

For this standard we have found neither explicit references to intercultural idiosyncrasies, nor references to VET. However, we know that approaches to programme errors or weaknesses and strengths can vary widely between cultures. For example, Germans are quoted as expressing their disagreement very bluntly and directly (‘You are wrong!’) and are very sparing with praise. The British, in contrast, ‘wrap up’ criticism or disagreement in polite phrases: ‘To a certain extent I agree with you, but I’m not totally convinced’, and may express agreement very strongly: ‘We see eye to eye on this affair’ (Bosewitz and Kleinschroth, 1997). An intercultural evaluation certainly demands ample knowledge and confidence in communicating strengths and weaknesses.

6.3.4. F4/P4: unbiased conduct and reporting

‘The evaluation should take into account the different views of the stakeholders concerning the evaluand and the evaluation findings. Similar to the entire evaluation process, the evaluation report should evidence the impartial position of the evaluation team. Value judgements should be made as unemotionally as possible.’

The nature of impartial conduct may differ between various nationalities and even between subcultures within a country. Evaluations in countries where VET institutions integrate social part-

ners almost automatically consider employer, union and public viewpoints so that they can be seen to be unbiased. In other cases, the status of a public organisation can indicate high dependence or a high level of independence. For example, evaluators who work full-time at a university are perceived to be less biased than those who work for a business consultancy or as freelancers, even if the reverse is true. In hierarchical organisations such as patriarchal companies or authorities, impartiality may be undesirable. This puts evaluators in a difficult position.

Culture affects preferences for the minimum necessary degree of consideration of various perspectives versus the maximum permissible, and their mode of representation. Public debate, which reveals clear differences of opinion, may either be inappropriate or second nature, depending on the culture (Smith and Jang, 2002).

In some cases it may even be very difficult to ascertain different viewpoints. Depending on the position on the ‘individualism – collectivism’ dimension (Smith and Jang, 2002), participants tend to present a more or less united front, particularly during group interviews. In other situations, a private conversation may well be perceived as an insinuation that group discussion does not permit the frankness desired. Choice of method can also encourage or hinder the disclosure of stakeholder perspectives.

Impartiality can be especially problematic when evaluators help develop the programme, formatively support its implementation and then describe and assess its results and effects ⁽⁸⁸⁾. This inevitably leads to role conflicts, challenging professional competence to the utmost. This conflict could, no doubt, be avoided by replacing the evaluation team between the formative and the summative stages. However, this would increase the cost of the evaluation. This standard thus places individuals in two or more incompatible roles in some evaluations ⁽⁸⁹⁾.

6.3.5. F5/P5: disclosure of findings

‘To the extent possible, all stakeholders should have access to the evaluation findings.’

This DeGEval standard focuses on informing the stakeholders. ‘If an evaluation should serve to

⁽⁸⁸⁾ This is typical of pilot projects run by the BIBB (Section 4.1).

⁽⁸⁹⁾ The SEVAL K6 standard Declaration of conflicts of interests is formulated ‘more realistically’ than the DeGEval standard.

improve, justify and boost comprehension of continuing training initiatives, the relevant parties should also have access to the evaluation investigations.’

Reischmann (2003, p. 256) goes a step further, referring to JC standard K6. He points out that if confidentiality does not dictate otherwise, it is sensible to disseminate the report, e.g. among interested colleagues, decision-makers, the mass media and academic journals. This is the way to reach a much broader audience, i.e. academics, politicians and the general public.

Publication of findings can trigger conflicts in vocational training pilot projects. Evaluators and academic research institutions are predominantly interested in publishing their findings because they largely influence their reputation in academic circles and/or on the evaluation market. Pilot project sponsors, in contrast, have little interest in, or are even opposed to, publication as they fear that the competition could benefit from their knowledge (Zimmer, 1998, p. 600). They are against any transfer of findings other than self-presentation as an innovative enterprise. According to Zimmer, evaluators also tend to not to favour transfer, as further pilot projects in other companies could lead to new contracts.

The consequences of this standard for evaluations conducted through private enterprises, the state or public institutions, i.e. institutions with tax advantages (such as foundations) must be adapted in various ways, as the phrase ‘[...] as far as possible’ indicates. In enterprises, ‘public’ basically means the entire company, and therefore encompasses management, staff and shareholders. The publicly financed sector targets a much wider audience, incorporating the mass media and citizens. If a particular evaluation is largely funded as a public-private partnership, its commissioning must contain optimal clarification to avoid subsequent disagreements and even legal action.

Public commissioners also have interests which must be protected, e.g. when a programme is still being developed and the evaluation commissioner discovers serious deficits at an early stage. Arrangements should be made for this eventuality.

This standard is closely related to the accuracy standards: ‘the publication of the research

methods, in particular of the identification assumptions underlying the derivation of a set of results, and on statements regarding the extent of any remaining uncertainty’ (Schmidt, 2001, p. 7) is important.

6.4. Commentary on the accuracy standards

‘The accuracy standards are intended to ensure that an evaluation produces and discloses valid and useful information and findings pertaining to the evaluation questions’ (DeGEval, 2002, p. 10).

The nine standards in this group can be broken down into four categories. The first two standards (G1/A1 and G2/A2) address the definition of the evaluand in context and demand a description of it. The next two standards (G3/A3 and G4/A4) demand identification of purpose, procedures and the information sources used in the evaluation. The next four standards refer to the actual processes for collecting, monitoring, evaluating and utilising data to draft conclusions. They define requirements for gathering and sifting information to reach findings. Standard G9/A9 imposes meta-evaluations as a method of evaluation quality assurance and improvement.

Literature on quality requirements for VET evaluations only touches on some aspects of the accuracy standards. The standards for empirical data gathering, in particular, match the criteria which apply to the quality of scientific investigations in general. Social science textbooks detail these criteria, possibly explaining why they are not addressed separately⁽⁹⁰⁾. However, we must remember that evaluations are not conducted exclusively by empiricists and that evaluation commissioners using the standards should be taught to recognise ‘good evaluation’ characteristics. Standard sets for these accuracy requirements have been issued by national trade organisations, academic societies and research promotion institutions like the German Research Committee (*Assurance of good scientific practice*)⁽⁹¹⁾.

As evaluations are a special application of empirical scientific methods, the quality criteria,

⁽⁹⁰⁾ See, for example, the references in the explanatory notes to the DeGEval Standards G5/A5 and G6/A6.

⁽⁹¹⁾ *ibid.*

which stem from basic research, must be adapted to the evaluand environment. This goes for VET and other evaluation applications.

Curiously, VET literature on accuracy standards focuses more on pilot training projects and macroeconomic evaluations than on business studies. The first two types prioritise generalisability of findings, whereas application in the third area often only concerns one company.

6.4.1. G1/A1: description of the evaluand

‘The evaluand should be described and documented clearly and accurately, so that it can be unequivocally identified.’

Description of the evaluand enhances understanding of results and findings and clarifies whether, and how far, they can be transferred to similar programmes. The entire programme or its parts and their relevant characteristics should be specified. A distinction can be made between:

- (a) concept: goals, content, didactic focus, duration, scope in hours of attendance;
- (b) input: number, gender, age and previous qualifications of participants and number and qualifications of trainers;

- (c) structure: sponsoring organisation, premises, financial outlay, teaching and study aids.

Alongside this informative function, results transferability can only be assessed if the evaluand is identifiable. This standard is closely linked to disclosure of findings (F5/P5), G3/A3 and G5/A5 and is also discussed there.

Users often do not value evaluation reports which lack such basic information.

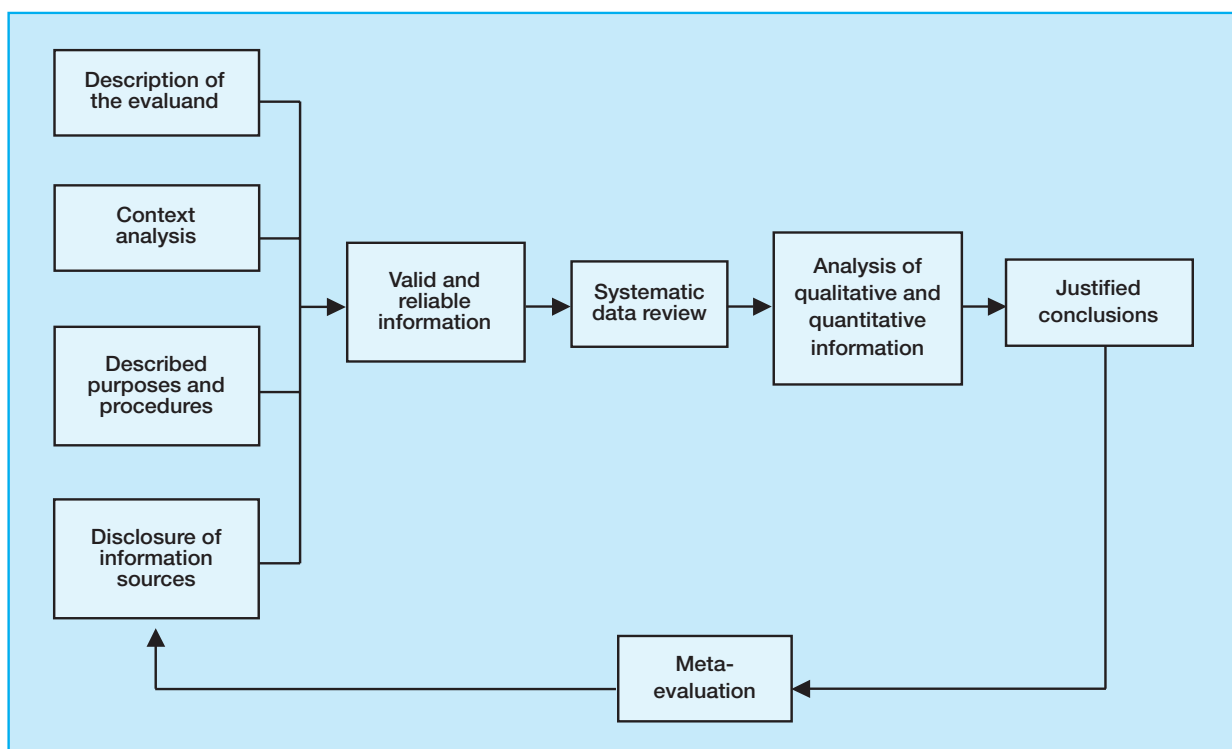
6.4.2. G2/A2: context analysis

‘The context of the evaluand should be examined and analysed in enough detail.’

Schmid (1996) notes that conventional evaluation research tends to neglect economics. Consideration of the context, however, is vital when inter-regional, intertemporal or international lessons are at stake. Turbin (2000) also states that VET systems are firmly rooted in their social context. If we are to learn from evaluations – best practices in other countries – the socioeconomic variables which affect these programmes and policies have to be identified. The ultimate goal is to assimilate the lessons in different environments.

Context also includes programme implementation conditions. Schmid calls for both programme

Figure 8: **Nine components which make evaluations accurate**



evaluation and goal-oriented evaluation procedures, describing them as ‘guidelines for international comparative research’ (Schmid, 1996, p. 205). His type of analytic strategy includes structural components of labour-market policy regimes and institutional components. Here he is implicitly referring to the organisational structure of political regimes, their responsiveness or implementation forms and their organisational efficiency (Schmid, 1996, p. 210).

These elements far exceed mere context description and could fall under evaluation purpose (N2/U2). This underlines the demand for context description.

According to Zimmer (1998) and Kaiser (1998), an evaluation of vocational training pilot projects naturally includes analysis of the social, economic, technical, occupational, cultural and educational environment, requirements and conditions for the project concerned and articulation of these circumstances.

Stakeholder information needs (A1/U1) and the evaluation purpose (N2/U2) dictate the scope and depth of programme context description.

6.4.3. G3/A3: described purposes and procedures

‘Object, purposes, questions and procedures of an evaluation, including the applied methods, should be accurately documented and described, so that they can be identified and assessed.’

The evaluation purposes should be specified and described both as a basic orientation in the detailed planning and in the evaluation report (N2/U2). The questions formulated at the beginning of the evaluation and the way they were adapted and extended should be recorded so that it is possible to judge whether the evaluation has answered them adequately (N4/U4). Timing, phases, methods applied, sampling and evaluation procedures should be presented. The description should also document any subsequent changes.

Although the DeGEval standards are constructed as maximum standards with scope for interpretation, as explained above, this standard specifies elements which no evaluation description should lack. Due to G3/A3’s univer-

sality, no VET adaptation needs exist, as in G4/A4 and G6/A6.

6.4.4. G4/A4: disclosure of information sources

‘The information sources used in the course of the evaluation should be documented in appropriate detail, so that the reliability and adequacy of the information can be assessed.’

This standard is also fundamental for assuring valid empirical practice. It should reduce the danger of ‘unscientific procedure’, particularly in research. Sources should be quoted precisely to guarantee intersubjective reliability. Standard G7/A7 is also relevant here. A mixture of methods, both qualitative and quantitative, can reduce the risk of erroneous procedure (Antoni, 1993) ⁽⁹²⁾.

Summative evaluations generally use quantitative methods, and formative evaluations usually employ qualitative methods, although it is virtually impossible to separate the two approaches strictly. Moreover, evaluators are often required to use both a formative and a summative procedure and apply the two methods appropriately. Antoni (1993) advocates an integrative approach. However, there is a danger that a comprehensive evaluation approach could lead to insufficient control and an increase in costs, rendering the evaluation impossible to implement. Data on general and vocational training and training and labour markets should be appropriately correlated. Discussion of the use of qualitative and quantitative information often reflects the conflict between micro- and macro-perspectives.

Incorporation of macroeconomic data when evaluating initial and continuing training is often requested, as public investment may be involved. Benefits to society as a whole and not just for the individual participants or a company and its profit or efficiency, are then of interest. Brüss (1997, p. 119 f.) argues against an obligation to link micro- and macro-perspectives in evaluations. If the programme budget constitutes only a tiny fraction of government spending, it is almost impossible to measure any macroeconomic effects of the programme. Furthermore, labour-market programmes on employment show that their influence is outweighed by other factors, such as general business trends.

⁽⁹²⁾ See more detailed discussion under G7/A7.

James and Roffe (2000, p. 13) point out that problems can arise between evaluators and commissioners if the latter have specific methods in mind. Evaluators must then justify themselves if they opt for less well-known methods, such as focus groups.

The literature shows there has been in-depth discussion of method selection and application for VET evaluations. Methodological issues often influence the choice of a specific evaluation model (Section 2.4).

We recommend referring to the *International handbook of labour market policy and evaluation*. This manual presents various methods and endorses them for certain investigations, e.g. experimental and non-experimental designs for evaluations of labour-market policy. These approaches are hotly debated in Europe ⁽⁹³⁾.

The goal-oriented evaluation model is designed to avoid the negative effects of one-dimensional impact assessment. They could be avoided by more complex analysis, including study of the socioeconomic context, monitoring and impact. Process-based and dialogue-oriented evaluation procedures are preferred (Schmid et al., 1996, Bangel et al., 2000). These authors urge marrying quantitative and qualitative procedures.

Empirical methods and tools should be properly tailored to evaluation purposes and the evaluand. Most authors claim neutrality in terms of the various research models. However, methods and tools dictate structure and are in no way neutral. The DeGEval G4/U4 standard commentary on analysis of qualitative and quantitative information demands that attention be paid to the validity of methods and their limitations. Kaiser (1998, p. 540) calls for elaboration of an evaluation concept and the publication of survey methods and data processing systems. We could go a step further and demand full disclosure for the evaluation models as well as for the methods employed.

6.4.5. G5/A5: valid and reliable information

'The data collection procedures should be chosen and developed and then applied in a way that ensures the reliability and validity of the data with regard to answering the evaluation questions. The technical criteria should be based on

the standards of quantitative and qualitative social research.'

Validity and reliability are fundamental prerequisites for empirical investigations. Numerous distinctions exist, e.g. between internal and external validity, content, criterion and construction validity, etc. These originated in quantitative research (e.g. testing procedures).

This is highly relevant for VET evaluations which use quantitative methods such as aptitude tests, personality inventories or standardised achievement tests for evaluation purposes. These gauges of quality are also essential for evaluation models which are based primarily on quantitative procedures. Schmidt (2000, p. 429) therefore advocates including an appropriate control group in labour-market policy evaluations. A convincing programme evaluation would hinge on this. Schmidt claims process analyses or before-and-after comparisons cannot replace this comparison situation. The literature speaks of the 'fundamental evaluation problem', as a counterfactual situation must often be postulated. Non-experimental procedures require educated estimates of what would have happened if trainees had not participated in the measure.

Commissioners often have input regarding method selection and choice of appropriate merit criteria, or they have preconceptions of what method to use. We can also assume that method selection also depends on the evaluator's speciality. Like Schmidt (2000), leading econometricians demand the construction of quasi-experiments or comparison situations. Evaluators with a teaching background may prefer to gather biographical data and narration from programme participants.

Different evaluation approaches entail differing methodological preferences. One approach adheres to the university research tradition and is often employed by academics. It chiefly relies on quantitative methods and indicators. Seyfried (1998) says this is too far removed from reality. Other approaches favour management methods even for evaluating training programmes. For example, the European Foundation for Quality Management process measures the quality of findings with either (monetary) benchmarks or participant statements. This engenders consider-

⁽⁹³⁾ Heckman and Smith (1996) or Nobel Lecture, Heckman (2001); see also Schmid (1996).

able validity problems (what is being measured: the learning and transfer results of the training or purely teacher or trainee attitudes?)

We recommend considering Cronbach's position⁽⁹⁴⁾. He describes evaluation as an art which, as such, differs fundamentally from science. He maintains that each evaluation involves an attempt to supply the commissioner and other interest groups with the maximum useful information for the given situation. Methodological standards, therefore, sometimes play a subordinate role in evaluation research. One would sometimes have to be content with a 'fair research design' and chiefly consider commissioner and stakeholder interests.

Undoubtedly, evaluations constantly have to compromise methodological quality in the face of tight deadlines and budgets, but failure to reach a certain minimum methodological level should be regarded as substandard and unacceptable.

These two stances were outlined to demonstrate briefly possible interpretations of quality requirements for evaluation methods. They represent a broad spectrum of opinions on (internal and external) validity and reliability and weigh them differently. Stufflebeam (2001) provides an overview of the various evaluation models, which also require different methods.

The standards demand valid information and this element of validity is not subdivided into internal or external validity. Various evaluation researchers⁽⁹⁵⁾ believe that external validity, i.e. the extent to which we can generalise findings, is a research issue and not an evaluation criterion. They say we must draw a clear line. Even if no explicit distinction is made between external and internal validity, the first sentence of G5/A5 suggests that it refers to internal validity, i.e. answering the evaluation questions, which will rarely refer to external validity. However, the second sentence of the standard speaks of criteria for quantitative and qualitative social research merit.

For internal validity, i.e. when effects appear which have not been caused by training measures, we must note the following factors: history, maturation, testing, instrumentation,

selection, mortality and the Hawthorne effect. Awareness of these potential effects is essential to avoiding them.

One final critical note on this standard: the merit criteria for validity and reliability are quantitative social research traditions. Some authors maintain that these criteria can be transferred, or at least adapted, to qualitative methods (Bortz and Döring, 2002, pp. 327-329). Others propose separate criteria for qualitative methods, such as trustworthiness instead of validity, dependability instead of reliability, and transferability instead of generalisability (Guba and Lincoln, 1989, pp. 233-251).

6.4.6. G6/A6: systematic data review

'The data collected, analysed and presented in the course of the evaluation should be systematically examined for possible errors.'

Collected facts and figures must be checked for accuracy. Pitfalls can occur in any phase of data gathering and evaluation and infallibility cannot be guaranteed. Wottawa and Thierau (1998) therefore recommend correcting project-related errors by means of organisational measures.

Professional standards dictate plausibility tests following data analysis. Plausibility tests involve identifying improbable data, e.g. by checking minimum/maximum values, compiling ratios (e.g. continuing training expenditure in euro per employee), calculating group averages, etc. Monitoring homogeneity (obtained from the variance) of the data actually obtained in relation to the overall sample can provide key indicators. In addition, Schmidt (2000) stresses stating all potential sources of error in an evaluation report.

6.4.7. G7/A7: analysis of qualitative and quantitative information

'Qualitative and quantitative information should be analysed in an appropriate, systematic way, so that the evaluation questions can be effectively answered.'

Elaboration of data evaluation plans before commencing the actual data processing is indispensable. Interpretation can then be tailored to the questions and assumptions which form the basis of the evaluation design. This also encourages

⁽⁹⁴⁾ Cronbach (1982) refers not only to VET but also to educational and social programmes which encompass VET.

⁽⁹⁵⁾ Fitz-Gibbon, among others, at the EES conference (EES 2002) expounded the opinion that monitoring external validity could not be the task of an evaluation.

targeted and efficient interpretation, which is essential for both quantitative and qualitative approaches.

Antoni (1993) writes that a combination of quantitative and qualitative procedures is often appropriate for meeting the situational requirements of evaluation problems in organisational and work psychology in particular. '[...] It is clear that qualitative and quantitative methods should be used in a complementary fashion in order to derive the highest value from research in this area,' observes Barrett (1998, p. 20) in one chapter on the relationship of quantitative and qualitative methods in continuing vocational training. He emphasises the importance of qualitative methods for enhancing quantitative approaches and vice versa. Gaude (1997) also underlines the significance of linking quantitative and qualitative procedures. Only quantitative procedures can demonstrate the effects of further training programmes on employment and income. Qualitative studies, in contrast, are necessary to explain why some programmes are more successful than others. They are also the only means of showing possible ways to improve programmes. Gaude deplores the mutual isolation of the various research disciplines, which means the more qualitative evaluations often lack information on the effects on income and employment, and qualitative evaluations are seldom conducted as part of quantitative evaluations.

6.4.8. G8/A8: justified conclusions

'The conclusions reached in the evaluation should be explicitly justified, so that the audiences can assess them.'

Conclusions condense the data gathered and their interpretation into findings, which may take the form of tenets, for example. This is a separate, essential task which the evaluator must perform. Reports which mainly present data, including diagrams, but which make no effort to draw conclusions from the summarised findings, are unacceptable.

However, it must also be possible to follow the argumentation of conclusions. Rieper (1997; p. 42) quotes a European expert examination of *ex post* evaluations which reveal that 'most reports neglected to explain the type of information on which their conclusions were based and how this information had been obtained.' This lack of transparency with regard to the methods applied makes

it difficult to assess the credibility and the potential use of the results and conclusions.

Disclosure of specific difficulties, survey methods and forms of data interpretation is particularly important for evaluating pilot projects (Kaiser, 1998, p. 540f). Some authors also believe that a summary of the findings at the end of the investigation should be accompanied by derived recommendations. Conclusions and recommendations are closely connected. 'To summarise, we can conclude the following from this measure on the basis of our evaluation [...] We deduce the following recommendations [...]' (Reischmann, 2003, p. 255).

The DeGEval standards do not specify that recommendations should be part of the evaluation or the report, as the evaluation model determines whether the evaluator is responsible for providing recommendations as well as drawing conclusions.

6.4.9. G9/A9: meta-evaluation

'The evaluation should be documented and archived appropriately, so that a meta-evaluation can be undertaken.'

Seyfried (1998) comments that very little transparent communication and discussion of methods and findings from VET evaluations can be found in Europe. Meta-evaluations could redress this. Lindley (1996) has developed a model for increasing evaluation transparency for European ESF projects which should also facilitate meta-evaluation. He proposes compiling a European evaluation database, which would list evaluations according to key indicators and record programme characteristics and their varying effects. He also recommends creating spatial typologies. Rieper (1997) also believes that comprehensive meta-evaluations are necessary to improve the quality of European Commission evaluations. The EU is a major commissioner.

Publication of evaluation reports could afford more opportunities to conduct meta-analyses as well as meta-evaluations. Disclosure of findings, as F5/P5 stipulates, should really be categorised under impact of the evaluation results. In contrast, the meta-analysis option is particularly interesting for those not affected, such as researchers, programme developers and government budgeters. This is the only way for evaluation findings to flow into strategic planning and decision-making processes (Fay, 1997, p. 113).

6.5. Proposals for expanding existing standards

Below we list gaps or ambiguities in the DeGEval standards that should be discussed at European level and clarified when further developing and adapting the evaluation standards.

6.5.1. Selection of the evaluation model

Numerous evaluation models are presented in the outline of evaluation standards and detailed in European literature. They differ considerably in their epistemological basis, their identification of values, their focus on specific elements of the evaluated programme (e.g. goals versus process versus effects), and many other issues. The reasons for selecting a particular model and its assumed strengths and limits are seldom discussed when the evaluation contract is processed.

All the well-known sets of standards fail to prescribe explicit disclosure of the selected evaluation model. Such a standard could further clarify the interaction between commissioners and contract recipients. It would also encourage more explicit presentation of evaluation theory and expose it to critical debate. In any case, evaluators should give their grounds for selecting a particular evaluation model (or combination of several models) and review them when the mission is accomplished.

6.5.2. Selection of suitable methods

Choice of method is often, although not always, closely linked to selection of the evaluation model. The DeGEval standards mention method selection frequently. In standard N4/U4, Information scope and selection, method choice focuses on the utility of the information gathered. Standard D1/F1, appropriate procedures, prioritises minimising inconvenience to the evaluand and the stakeholders in relation to the expected benefit of the evaluation. The explanatory notes to this standard point out that 'the most conclusive methods from a scientific point of view are often unsuitable because they are too laborious or ethically unacceptable in the situation concerned. The evaluation team should clarify advantages and disadvantages and justify the relevance of the chosen procedure.'

Some sources scrutinise methodological aspects. Surprisingly, the DeGEval standards

feature no separate standard on investigation design choice and thus method justification. Methods should encourage optimal response to the evaluation questions. The MEANS handbooks contain short guides to selecting various methods at different points in the evaluation (prospective versus retrospective analysis) and for different types of evaluands (e.g. overall programme evaluation versus in-depth evaluation tools). (European Commission, 1999b, Vol. 3, p. 219).

Method selection often involves making and justifying a decision on control groups or other suitable survey designs. The literature frequently insists that various levels, such as micro- and macro-evaluation, must be dovetailed if evaluation is to be meaningful. It also focuses on the problems of selecting and linking quantitative and qualitative survey methods. Here we discern a gap which expanding the existing standards or formulating a new standard could close.

Selection of the right evaluation methodology could be crucial. It is a vital condition for evaluation success. Many evaluation methods exist. Not every evaluation method is suitable for every evaluation purpose. The optimal solution depends on the questions and the solutions sought (evaluation purpose). 'It is important to choose the right methodology for evaluation, and there is a broad range to choose from. The fact that there are so many different approaches in use really reflects the view that no single methodology can be universally applied. The optimum choice depends on the questions and solutions that are sought.' (James and Roffe, 2000, p. 17)

The MEANS handbooks consider method selection in relation to the evaluation team profile and to assessment of the quality of an evaluation bid. European Commission (1999b), Vol. 1, pp. 82-83. They point out that choosing whether to award a future evaluation contract to a business consultant or university researcher also constitutes a decision for or against a certain approach. They suggest setting a budget, rather than stipulating a method, ideally in invitations to tender, and then selecting the team with the most interesting methodological proposal. The commissioner must then judge bid quality in a subsequent step. The method or methods must be the best to answer the prescribed questions.

6.5.3. Explicit reference to evaluation of training programmes

Standard P1 from the JC standards (service orientation support) does not feature in the German and Swiss evaluation standards as it refers explicitly to training programme evaluation, whereas they are intended to be universally applicable. Standard JC-P1 reads as follows. ‘Evaluations should be designed to assist organisations to address and effectively serve the needs of the full range of targeted participants.’ It adds that evaluations should also play a supporting role in ensuring that education and training goals are appropriate and that sufficient attention is paid to learner development, that promised services are rendered and that non-beneficial or even harmful programmes are abandoned. In this way evaluations should contribute towards making projects accountable to stakeholders and society. Evaluations in the VET sector should basically be designed to serve the interests of current or future learners.

Evaluators, commissioners and politicians must look beyond the short and medium-term interests of programme organisers and sponsoring organisations and also focus on the development of the educational system and its interaction with society.

An additional VET standard corresponding to JC-P1 would be feasible. It could be based on the guidelines to this JC standard, which include the following ⁽⁹⁶⁾:

- (a) ‘evaluations should be planned which foster the quality of programmes for education, initial and continuing training.’;
- (b) ‘evaluations should serve to identify intended and unintended effects of the programme on the learners.’;
- (c) ‘teaching and learning processes should be disrupted as little as possible, but an effort should be made to realise the evaluation project.’

JC-P1 content and comments could be highly relevant to VET evaluations in Europe.

⁽⁹⁶⁾ Guidelines A, D and H.

7. Summary and outlook

7.1. Objectives, questions and method of the study

The objective of the report is to reflect the transferability of evaluation standards in the European VET context. The following initial questions are considered:

Does the terminology of the standards match concepts in the area of European initial and continuing vocational training? Are any standards not applicable in the context of initial and continuing vocational training? Do European evaluation experts understand and accept the key concepts (e.g. definition of evaluation, differentiation between formative and summative evaluation, purpose of evaluation, etc.) conveyed? Are there any specific national differences which should be considered in defining standards? The standards of the DeGEval (2002) form a reference point for the analysis. Other relevant standards are presented and reflections on intercultural transferability and applicability to the VET evaluand are made. In further discussion the opinions of experts are included. This occurs first in documented events on the standards attended by vocational training experts in Germany and Austria. Second, evaluation experts in widely divergent European countries are sent a questionnaire. Finally, the DeGEval standards are also debated in commentaries and in the formulation of criteria in recent European literature on VET evaluations.

7.2. Results and conclusions

7.2.1. Standards for programme evaluation

The background, evolution, and constitution of DeGEval's evaluation standards are described in some detail. The DeGEval standards draw heavily on the constitution and content of the most widely known standards of the JC. In 1981 the committee first published the *Standards for evaluation of educational programs, projects and*

materials (JC, 1981) and issued a revised edition, *The program evaluation standards*, in 1994. DeGEval standards consist of standards for evaluation assigned to four different groups with explanatory notes and annexes (DeGEval, 2002), the English translation of which is attached to this report. Evaluations should thus demonstrate the following four basic attributes: utility, feasibility, propriety and accuracy. It is presumed that an evaluation will simultaneously take account of all four criteria to fulfil specialist and professional requirements. The 25 standards are divided into these four categories.

7.2.2. Transferability of standards

To summarise, we can say that there are no discrepancies or contradictions between the various European evaluation standards presented here. Their respective evolution, emphases and differentiation bear witness to different approaches. The development of standards in some European countries, such as Germany and Switzerland, draws on US evaluation standards in their constitution and contents. Other countries, such as France and Finland, attempt to create or adopt their own. In France, for example, the aspect of social utility is more intensely discussed, while in UK attention is primarily given to the coordination processes between the various interest groups. Differences are also apparent with regard to the varied portrayals. Some sets of standards are more generally formulated – such as Finnish ethical guideline standards – while other sets are very concrete and prescriptive, such as those governing the readability of reports. Many of the standards considered contain the principal components of the DeGEval standards. Equally, the European Commission's MEANS criteria and the guidelines of the International Labour Office (ILO) show some similarities with DeGEval standards or their American bases.

The original US standards were initially used in education and were then applied to programme evaluations in other policy areas, which in turn influenced their content. Since the standards

originated in education and are supposed to be applicable to all policy areas, the initial presumption is that they are also valid in VET. Furthermore, seven illustrative examples from JC Standards are drawn from in-company vocational training, and are therefore directly applicable to VET.

7.2.3. Results from group discussion on the applicability of DeGEval standards to vocational training

In general the standards for evaluation have been well received by evaluation and/or VET experts who participated in investigations within the framework of the discussion. Neither the German nor the Austrian experts have any reservations as to the applicability of evaluation standards to the field of vocational and continuing training. No doubts were expressed as to the transferability of the standards to vocational training as an evaluation with specific institutional arrangements (for example, the dual system of vocational training in Germany). The experts do not propose a specific adaptation, although they would like to see certain standards illustrated by examples from initial and continuing vocational training.

In line with the publications of the Joint Committee, there is a call to have standards complemented by extensive explanatory notes with guidelines and practical examples from VET. It would be a great advantage for these to include explanations of concepts such as the difference between stakeholders, addressees and other users. Furthermore, workshop participants, particularly academics from universities and public and independent research institutes, have drawn attention to the stark conflict between the standard group's utility and accuracy. Expectations of immediately utilisable results and the demands of empirical social and economic research for such qualities as the validity of instruments and the reliability of data compilation, are often in direct competition.

Some participants, who have been working for years within a particular discipline or theoretical tradition, have expressed initial concerns that their methodology might not be adequately covered by the evaluation standards. A general desire for maximum standards has been accompanied by ambivalence to them. On the one hand, participants appreciate the advantage that maximum standards can refer to many different approaches and types of evaluations and impact

investigations. On the other hand, representatives of certain schools of thought bemoan the lack of obligatory minimum standards. Furthermore, participants want more emphasis on the fact that the standards do not prescribe a given evaluation approach. The explanatory notes to the standards already state that there are 'numerous different approaches to professional evaluation' and that these contrast starkly depending on epistemological approach, discipline and professional ethics. The pluralistic foundation of the standards is sometimes, however not immediately, clear to experts the first time they read the text. They often worry that the standards will have a restrictive effect on the approach they advocate, or even exclude it entirely.

One important point for further research should be the question of whether other European countries are aware of self-evaluation approaches to VET evaluation or whether they have their own interpretation of external or internal independent evaluation.

7.2.4. Survey of evaluation experts in Europe

The experts interviewed generally have a positive attitude towards standards for evaluation. None of the respondents feel that standards do not matter or are unnecessary or even harmful; and they express a preference for maximum standards.

The best-known sets of standards are the *US joint committee standards for evaluation* and the *American guidelines for evaluators*. The vast majority of the respondents named a minimum of one set of standards with which they are at least familiar.

The main benefits of evaluation standards named in critical discussion are improvement in the quality of evaluations and the opportunity to make evaluators' work more legitimate and transparent. However, they fear that utilisation of the standards could restrict the plurality and flexibility of evaluations in theory and in practice, or that standards could be applied too rigidly. The majority of respondents mention no preferable alternative to the standards.

Respondents named involvement of all stakeholders, transparency and use of a wide variety of suitable methods as the most important hallmarks of evaluation standards.

7.2.5. Reflections on VET evaluation standards literature

Not all standards are equally applicable to every evaluation project. This also goes for VET evaluations, of course. Nevertheless, the validity of each individual DeGEval standard has been confirmed in various VET contexts. The literature studied contains concrete quality requirements, advice and guidance on using evaluations which resemble the individual DeGEval standards. We can thus illustrate the individual standards in terms of the VET evaluand and its characteristics. Moreover, we can refine some of the standards further.

The text, therefore, features a brief introduction to each group of standards, followed by a commentary on individual standards. This commentary is partly illustrative and descriptive, and partly more reflective, depending on the conflict potential which each VET standard contains. At the same time, it is noteworthy that the utility, feasibility, and propriety standard groups yielded many more points of reference for VET evaluation than do those of accuracy, which repeatedly formulate universal demands on empirical investigations.

The polarity between the groups of standards relating to accuracy and utility proves to be the cause of a lasting and irrevocable clash. On occasions, expectations of immediately utilisable results and the demands of empirical social and economic research for such qualities as validity and reliability are scarcely reconcilable, so that either one or the other must make sacrifices.

7.3. Outlook

The following contains proposed tenets for the utilisation and elaboration of evaluation standards in European initial and continuing training.

European and national organisations working in VET should determine a set of standards by a given deadline to provide orientation and guidelines for professional VET evaluations they commission.

The selection and prescription of such a set of standards should be undertaken through dialogue between European evaluation societies and supplemented by academic specialist and professional associations operating particularly in the field of VET. It may be advisable to allow

associations, in particular the European Evaluation Society, to take the initiative.

The text accompanying the evaluation standards should emphasise that, in the light of national and disciplinary peculiarities, evaluation theory and practice evidences divergent traditions and models and that appropriate adaptation of standards is possible and desirable.

The duty of evaluation to the 'common good' may be addressed in the context of national evaluational tradition, but should, however, be set at a European level in cases where broader consensus exists.

The application of evaluation standards throughout Europe demands a suitable degree of intercultural competence on the part of evaluators and evaluation commissioners, which is to be promoted by appropriate training and processes of systematic reappraisal (e.g. in European evaluation journals and international congresses).

Standards are to be maximum standards. The description should be as plain as possible and cite the ideal that an evaluation is to strive towards in the respective categories for it to be judged high quality. Such maximum standards offer clear orientation, but also leave sufficient scope for flexibility, national and local adaptation.

Attention is to be drawn to the inappropriateness of rigid application of evaluation standards and to continuing incompatibility between individual standards, particularly between those of accuracy and utility.

Publications on evaluation standards in Europe are to contain key definitions for concepts such as evaluation, evaluation model, evaluation purpose, evaluation questions, formative evaluation, process evaluation, etc. A multilingual glossary could improve cooperation between evaluators working within European programmes and policy-making.

Sets of standards are to contain evaluation standards that generally apply to evaluation and therefore also to VET evaluation. Accompanying material must be made available for VET. This should offer illustrative examples from conducted VET evaluations for as many relevant system levels as possible (i.e. self-study, companies/schools, associations of learning locations, communities and regions, national and pan-European VET programme). This is essential to demonstrate the validity of evaluation standards to all VET system

levels and to minimise existent reservations among professionals who are unfamiliar with standards.

Because of the present gap in research, meta-evaluations are to review systematically whether evaluation standards are applicable and appropriate to VET evaluations of national programmes or to EU policies, and which additions are necessary, particularly to supplementary materials, to steer and evaluate evaluations.

A separate, general evaluation standard should be formulated which calls on those responsible for evaluations to explain the model or models used for a given evaluation and to justify its/their suitability to the evaluation in question. Such a call for disclosure and justification might support the propagation of evaluation models, the mooting of their strengths and weaknesses and the culture of meta-evaluation.

Formulation of an additional VET-specific standard is proposed. This standard, Quality orientation support in vocational training, could read as follows: 'Evaluations should assist VET policy-makers and programme managers to meet quality requirements within the vocational training sector (VET standards). These particularly include standards which require evaluations to consider the needs of target groups, social partners and society, have a scientifically founded theoretical and teaching concept, help shape the structure and organisation of political education and help

manage educational processes and ensure the profitability of VET activities.' The explanatory notes on this standard should mention well-known, recognised VET standards and point the way to the most important sources.

Since teaching personnel in VET in particular always question the tool of self-evaluation, evaluation standards should specify that this is primarily relevant to internal and external investigations undertaken piecemeal by specialists. Self-evaluation in the field of education should employ a set of standards oriented towards general standards and adapted to those ends.

As the professionalisation of evaluation is still fairly recent in the majority of Member States, further investigations, founded on a broad empirical usage of data, are particularly necessary to clarify matters bearing on the compatibility of culturally sensitive individual standards and the role of self-evaluation in VET. We consider workshops and conferences related to data collection, as used in these studies, especially useful to these ends.

This report appraises standards for programme evaluation. In the future, further evaluation standards such as Personnel evaluation standards and the Student evaluation standards are to be analysed in relation to VET and their transferability across Europe and their intercompatibility examined.

List of abbreviations

Abbreviations of individual standards used in text

Nn/Un (*Nützlichkeit/Utility*), Dn/Fn (*Durchführbarkeit/Feasibility*), Fn/Pn (*Fairness/Propriety*) and Gn/An (*Genauigkeit/Accuracy*) refer to DeGEval standards (2001). The capital letter to the left of the slash signifies the German original designation of the group of standards, the capital letter to the right of the slash the English translation.

The Arabic numeral to the right of each capital letter indicates the individual standard in the order in which it is listed in the appropriate group of standards (e.g. N2/U2, *Klärung der Evaluationszwecke/Clarification of the purposes of the evaluation*).

A standard from Group U, F, P or A of the Joint Committee standards (1994) is addressed by prefixing 'JC', e.g. JC-A7 – Systematic Information.

BIBB	German Federal Institute for Vocational Training
BifEb	Austrian Federal Institute for Adult Education
DeGEval	<i>Deutsche Gesellschaft für Evaluation</i> [German Evaluation Society]
EES	European Evaluation Society
ISO	International Organisation for Standardisation
JC	Joint Committee on Standards for Educational Evaluation
SEVAL	<i>Schweizerische Evaluationsgesellschaft</i> [Swiss Evaluation Society]
SFE	<i>Société Française de l'Évaluation</i> [French Evaluation Society]

Annex 1: transformation table

<i>Deutsche Gesellschaft für Evaluation</i>		Joint Committee on Standards (US)	
U1	Stakeholder identification	U1	Stakeholder identification
U2	Clarification of the purposes of the evaluation		missing
U3	Evaluator credibility and competence	U2	Evaluator credibility
U4	Information scope and selection	U3	Information scope and selection
U5	Transparency of values	U4	Values identification
U6	Report comprehensiveness and clarity	U5	Report clarity
U7	Evaluation timeliness	U6	Report timeliness and dissemination
U8	Evaluation utilisation and use	U7	Evaluation impact
F1	Appropriate procedures	F1	Practical procedures
F2	Diplomatic conduct	F2	Political viability
F3	Evaluation efficiency	F3	Cost effectiveness
	inapplicable	P1	Service orientation
P1	Formal agreements	P2	Formal agreements
P2	Protection of individual rights	P3	Rights of human subjects
		P4	Human interactions
P3	Complete and fair investigation	P5	Complete and fair assessment
P5	Disclosure of findings	P6	Disclosure of findings
	in P4 unbiased conduct and reporting	P7	Conflict of interest
	in F3 evaluation efficiency	P8	Fiscal responsibility
A1	Description of the evaluand	A1	Program documentation
A2	Context analysis	A2	Context analysis
A3	Described purposes and procedures	A3	Described purposes and procedures
A4	Disclosure of information sources	A4	Defensible information sources
A5	Valid and reliable information	A5	Valid information
		A6	Reliable information
A6	Systematic data review	A7	Systematic information
A7	Analysis of qualitative and quantitative information	A8	Analysis of quantitative information
		A9	Analysis of qualitative information
A8	Justified conclusions	A10	Justified conclusions
A4	Unbiased conduct and reporting	A11	Impartial reporting
A9	Meta-evaluation	A12	Meta-evaluation

Annex 2: questionnaire

Quality requirements for evaluations in vocational education and training (VET)

Dear

we would like to invite you to participate in a pilot study on quality in evaluation. Please answer our short questionnaire. We would need it back at least until

We contact you as one of about 30 experts in evaluation and/or VET from all European countries. We have got your name and address from, who recommended to contact you.

Aim of the survey

We should appreciate your answer to the question, whether or not VET evaluations in Europe need a professional codified framework for securing and enhancing the quality of evaluation practice. We also would like to ask for your advice: which values and demands should be considered in such a framework?

This study is commissioned by the European Centre for the Development of Vocational Training (Cedefop), an agency of the EU. The results of this study will be included into the third Cedefop research report entitled *Research on evaluation and impact of vocational education and training* which will be published in 2004. The title of our paper will be *Ethical and normative standards for evaluation practices*.

This is a pilot study!

We just started this pilot study to bring more clearness into an emerging field: the evaluation of VET measures and programmes in European countries. VET evaluation as a theme across the European countries is just in its prime and we consider to need an open dialogue to promote it! You can read more of the background of this pilot study in the attached description.

What is your investment? How to send back your answers?

It takes around 10 minutes to fill in the following questionnaire. If you are very short in time please answer all closed questions and skip one or another open ended question which may not be so important for you. If you want to comment some questions in more detail we would appreciate it.

Please print out the document, fill it in by hand and fax it back to us (+49 221 4248072). If you would like the document by fax please let us know (+49 221 4248071).

What do we offer for your active participation?

We will compile a documentation and summary of the answers on this questionnaire and will send this material and our theses/conclusions to you in autumn this year.

We would appreciate your feedback on our conclusions but this is really voluntary!

End 2002, we will post you an electronic version of the survey report and ask you whether or not you want to be mentioned as participant of the pilot study.

You can find the following files as attachments:

- (a) our questionnaire as a Word-file;
- (b) our questionnaire as a pdf-file;
- (c) a short description of the pilot study.

Thank you in advance for your kind cooperation.

Wolfgang Beywl

Sandra Speer

Univation – Institute for evaluation
Zuelpicher Str. 58
D – 50674 Koeln
Tel: +49 221 424 8071
Fax: +49 221 424 8072

Quality requirements for evaluations in vocational education and training (VET)

About this study

This study is commissioned by the European Centre for the Development of Vocational Training (Cedefop), an agency of the EU. The results of this study will be included into the third Cedefop research report entitled *Research on evaluation and impact of vocational education and training* which will be published in 2004. The title of our paper will be *Ethical and normative standards for evaluation practices*. In this study we will discuss important issues of applying standards to evaluations of VET measures and programmes.

What do we offer for your active participation?

We will compile a documentation and summary of the answers on this questionnaire and will send this material and our theses/conclusions to you in autumn this year. End 2002 we will post you an electronic version of the survey report and ask you whether or not you want to be mentioned as participant of the pilot study in this document.

What about confidentiality?

We will ensure full confidentiality of all information you give us and handle them anonymously. In the documentation we will only mention the country the respondent refers to (see question 3), and no names. After we have finalised the final report, we will send it to you and ask whether or not you want to be included in the expert list which will be added to the report. By doing so we want to enable you to express your considerations and arguments plus your emerging ideas and issues in an open manner.

Any questions/reservations?

Please do not hesitate to contact us by e-mail (cedefop@univation.org) or phone (+49 221 424 8071). We will answer your questions immediately and phone back if you wish so (in this case please attach your phone number).

Questionnaire

(please mark the box belonging to the fitting answer ☐)

- 1) Your primary position in/to evaluation (*choose one alternative*).
 - ☐ Client/sponsor /commissioner
 - ☐ Evaluator
 - ☐ Programme director/ programme staff
 - ☐ Other:(Please specify)
- 2) What is your main professional background? (*choose one alternative*)
 - ☐ Economics
 - ☐ Social and political sciences
 - ☐ Natural sciences
 - ☐ Liberal arts incl. pedagogics
 - ☐ Engineering
 - ☐ Other:(Please specify)
- 3) The national professional culture you mostly identify with (*this might be the country you have been educated/studied, the country you work in normally/at present, or it might be or not your nationality in your passport*).
 International country code:

- 8) How familiar are you with the following sets of standards/guidelines for evaluation? (Please mark one alternative within every row)

No		Very familiar	Quite familiar	Know a little bit	Do not know
1	Joint committee standards for evaluation (USA 1994) English: http://www.eval.org/EvaluationDocuments/progeval.html	☐	☐	☐	☐
2	Guidelines for evaluators (American evaluation association, 1994) English: http://www.eval.org/EvaluationDocuments/aeaprin6.html	☐	☐	☐	☐
3	The Means collection, European Communities, Directorate General XVI. Luxembourg, 1999. (Not available on the Internet)	☐	☐	☐	☐
4	Swiss evaluation society (2001) German: http://www.seval.ch/deutsch/stad/stad1.htm French: http://www.seval.ch/franz/staf/staf1.html	☐	☐	☐	☐
5	German evaluation society (2001) German: http://www.degeval.de/standards/standards.htm English: http://www.degeval.de/standards/Standards_engl.pdf	☐	☐	☐	☐
6	Best practice guidelines for evaluation of the OECD English and French: http://www.oecd.org/home [search]	☐	☐	☐	☐
7	Other:	☐	☐	☐	☐
8	Other:	☐	☐	☐	☐

- 9) Pro's for standards for VET evaluation.
Please write down some arguments (if any) which call for a more intensive use of standards for evaluation in VET evaluation.

.....

- 10) Con's against Standards of VET evaluation.
Please write down some arguments (if any) which speak against a more intensive use of standards for evaluation in VET evaluation.

.....

- 11) Are there essential omissions?
Please indicate essential omissions (what lacks?) in the standard set(s) you know which should be supplemented for VET evaluations, or indicate important demands/aspects a set of standards should include.

.....

- 12) Better alternative?
Is there some tool/regulation which suits better than standards for evaluation to enhance/secure quality of VET evaluations?

.....

- 13) Basic values which should be included in regulations for VET evaluations.
We are looking for basic values you would associate with 'good' VET evaluation. Please name one to five attributes which are essential for VET evaluation quality.

.....

14) Address for contact

Please state your name, phone number and e-mail address, so that we can contact you, if you have any question.

.....
.....

15) A second respondent you propose

Maybe you have an idea to whom else from your country or elsewhere the questionnaire should be sent. If you like, please state his/her name and e-mail address.

.....
.....

We would like to thank you for your kind cooperation.

Wolfgang Beywl

Sandra Speer

Annex 3: list of experts answering the e-mail survey

(12 out of 19 persons agreed to publish their names and addresses)

Name, surname Function in VET/evaluation	Institution, city, country, e-mail	Other functions in VET/evaluation
Barbier Jean Claude Research Director	<i>Centre d'Études de l'Emploi</i> Noisy-le-Grand – France Jean-Claude.Barbier@mail.enpc.fr	Evaluation of public policies
Bjørnkilde Thomas Manager	PLS Rambøll Management A/S Copenhagen – Denmark thomas.bjornkilde@pls-ramboll.com.	
Field Jane Consultant, specialising in evaluation and LLL	Education and Development Whitehead Co Antrim Northern Ireland jane@educationanddevelopment.co.uk	Author of Evaluating Community Development Projects; NIACE, March 2003
Franz Hans-Werner Researcher, consultant, manager	<i>Sozialforschungsstelle Dortmund</i> <i>Landesinstitut</i> Labour-related research and advice Dortmund – Germany franz@sfs-dortmund.de	VET, CVET, TQM, EFQM; several books and articles on the subject
Hartkamp Jannes Researcher	DESAN Research Solutions Amsterdam – The Netherlands hartkamp@desan.nl	Main fields: VET, transition from education to work, metadata standards
Kirsch Jean-Louis Researcher	<i>Centre d'études et de recherches sur les qualifications</i> Centre for research on education, training and employment Marseille – France jlkirsch@cereq.fr	Statistics, accompaniment of actions in the field of training and employment
Nicaise Ides	HIVA (Higher Institute for Labour Studies) and Dept of Education University of Leuven – Belgium ides.nicaise@hiva.kuleuven.ac.be	
Nurmi Johanna Senior Adviser	Finnish Ministry of Finance Public Management Department Valtioneuvosto – Finland johanna.nurmi@vm.fi	Secretary of the Finnish Evaluation Society (FES)
Rouland Olivier Administrator	DG Budget – Evaluation Unite Brussels – Belgium Olivier.Rouland@cec.eu.int	
Schiefer Ulrich	ISCTE – Higher Institute for Labour and Business Studies Lisbon – Portugal schiefer@iscte.pt	Board member of the European Evaluation Society,
Smid Gerhard Programme manager	Interuniversity Centre for Development in Organisation and Change Management Utrecht – The Netherlands Smid@sioo.nl	
Vedung Evert Evaluation teacher	Uppsala University Institute for Housing and Urban Research – IBF Gävle – Sweden Department of Government Uppsala – Sweden evert.vedung@ibf.uu.se	

References

- Antoni, C. H. Evaluationsforschung in der Arbeits- und Organisationspsychologie. In: Bungart, W.; Herrmann, T. (eds) *Arbeits- und Organisationspsychologie im Spannungsfeld zwischen Grundlagenorientierung und Anwendung*. Bern et al.: Huber, 1993, p. 309-337.
- Auer, P.; Kruppe, Th. Monitoring of labour market policy in EU Member States. In: Schmid, G. et al. (eds) *International handbook of labour market policy and evaluation*. Cheltenham: Edward Elgar, 1996, p. 899-922.
- Bangel, B. et al. Arbeitsmarktpolitik. In: Stockmann, R. (ed.) *Evaluationsforschung, Sozialwissenschaftliche Evaluationsforschung 1*. Opladen: Leske und Budrich, 2000, p. 309-341.
- Barrett, A. Methodological Introduction. In: Elson-Rogers, S. (ed.) *Approaches and obstacles to the evaluation of investment in continuing vocational training: discussion and case studies from six Member States of the European Union*. Luxembourg: Office for Official Publications of the European Communities, 1998, p. 18-24 (Cedefop Panorama Series, 5078).
- Beywl, W. *Zur Weiterentwicklung der Evaluationsmethodologie*. Frankfurt: Lang, 1988.
- Beywl, W.; Schobert, B. *Evaluation – Controlling – Qualitätsmanagement in der betrieblichen Weiterbildung: kommentierte Auswahlbiographie*. Bielefeld: Bertelsmann, 2000.
- Beywl, W.; Taut, S. Standards: Aktuelle Strategie zur Qualitätsentwicklung in der Evaluation. *DIW Vierteljahresheft*, 2000, No 3, p. 358-370.
- Beywl, W.; Widmer, Th. Die 'Standards' im Vergleich mit weiteren Regelwerken zur Qualität fachlicher Leistungserstellung. In: Sanders, J. R. (ed.) *Handbuch der Evaluationsstandards*. Opladen: Leske und Budrich, 2000, p. 259-295.
- Beywl, W.; Speer, S.; Kehr, J. *Wirkungsorientierte Evaluation in der Armuts- und Reichtumsberichterstattung – Eine Perspektivstudie*. Bonn: BMGS (ed.), 2003 (in print).
- Bezzi, C. Claudio Bezzi's statements about evaluation standards. In: *The 2002 EES Conference – Three movements in contemporary evaluation: learning, theory and evidence*. Sevilla, 10-12 October 2002 (Evidence – Roundtable EV SR 4: Do we need European evaluation standards?).
- BIBB – Bundesinstitut für Berufsbildung et al. (eds) *A European comparison of controlling in corporate continuing training*. Bielefeld: Bertelsmann, 2001.
- Björklund, A.; Regnér, H. Experimental evaluation of European labour market policy. In: Schmid, G. et al. (eds) *International handbook of labour market policy and evaluation*. Cheltenham: Edward Elgar, 1996.
- Bortz, J.; Döring, N. *Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler* (Third edition). Heidelberg: Springer, 2002.
- Bosewitz, R.; Kleinschroth, R. *Getting through at meetings. Business English für Konferenzen und Präsentationen*. Reinbek: Rowohlt, 1997.
- Brüss, K. Approaches to evaluate the effects of labour market policies in Germany, and in particular those co-financed by the European Social Fund (ESF). In: *Evaluation of European training, employment and human resource programmes*. Luxembourg: Office for Official Publications of the European Communities, 1997, p. 115-123 (Cedefop Panorama, 5062).
- Bustelo Ruesta, M. Deontología de la evaluación: el modelo de los códigos éticos anglosajones. *Gestión y Análisis de Política Pública*, 1998, No 11-12, p. 141-156.
- Butz, M. Evaluationsverfahren in der betrieblichen Weiterbildung im IT- und TK-Bereich. In: Heinrich, L.J.; Häntschel, I. (eds) *Evaluation und Evaluationsforschung in der Wirtschaftsinformatik*. Munich; Vienna: Oldenbourg, 2000, p. 423-438.
- Cedefop. *Ethical and Normative Standards for Evaluation Practices*. Thessaloniki, 2001 (Project 0730 [MT/PDE RR3-13]).
- Chen, H. *Theory-driven evaluation*. London: Sage, 1990.
- Cronbach, L. J. *Designing evaluations of educational and social programs*. San Francisco: Jossey-Bass, 1982.
- Danida – Danish Agency for Development and Aid. *Evaluation Guidelines* – February 1999 (2nd

- edition, revised 2001) Available from Internet: <http://www.um.dk/danida/evalueringsrapporter/eval-gui/index.asp> [cited 24.10.2003].
- DeGEval – Deutsche Gesellschaft für Evaluation (ed.) *Standards für Evaluation*. Cologne: Eigenverlag, 2002.
- EES – European Evaluation Society. *The 2002 EES Conference – Three movements in contemporary evaluation: learning, theory and evidence*. Sevilla, 10-12 October 2002. Available from Internet: http://www.europeanevaluation.org/general/ees_conferences.htm [cited 13.10.2003].
- ETF – European Training Foundation. *Development of standards in vocational education and training*. Luxembourg: Office for Official Publications of the European Communities, 1999, Vol. 1.
- European Commission. *Evaluating EU expenditure programmes: a guide. Ex post and intermediate evaluation*. Directorate-General XIX-Budgets, January 1997. Available from Internet: http://europa.eu.int/comm/budget/evaluation/pdf/guide_en.pdf [cited 24.10.2003].
- European Commission. *Guidelines for systems of monitoring and evaluation of ESF assistance in the period 2000-2006*. Luxembourg: Office for Official Publications of the European Communities, 1999a. Available from Internet: http://europa.eu.int/comm/employment_social/esf2000/guidelines/evaluation/en.pdf [cited 24.10.2003].
- European Commission. *The MEANS Collection*. Luxembourg: Office for Official Publications of the European Communities, 1999b, Vol. 1-6.
- European Commission. DG for Education and Culture and DG for Employment and Social Affairs. *Making a European area of lifelong learning*. Luxembourg: Office for Official Publications of the European Communities, 2002.
- Fay, R. G. What can we learn from evaluations of active labour market policies undertaken in OECD countries? The case of training. In: *Evaluation of European training, employment and human resource programmes*. Luxembourg: Office for Official Publications of the European Communities, 1997, p. 101-113 (Cedefop Panorama, 5062).
- Fetterman, D. M. *Foundations of empowerment evaluation*. Thousand Oaks: Sage Publications: 2000.
- Field, J. *Promoting added value through the evaluation of training (PAVE) Handbook*. University of Plymouth, 1998a.
- Field, J. *Promoting added value through the evaluation of training (PAVE) Evaluation Resource Pack*. University of Plymouth, 1998b.
- Field, J. Promoting added value through the evaluation of training-PAVE. In: Künzel, K. (ed.) *Evaluation der Weiterbildung*. Internationales Jahrbuch der Erwachsenenbildung. Cologne: Böhlau, 1999, Vol. 27, p. 215-229.
- FES – Finnish Evaluation Society, 2002. Available from Internet: http://www.finnishevaluationso-society.net/index_en.php [cited 24.10.2003].
- Finné, S. Guidelines for the intermediate monitoring and evaluation of Structural Fund operations. In: *Evaluation of European training, employment and human resource programmes*. Luxembourg: Office for Official Publications of the European Communities, 1997, p. 47-50 (Cedefop Panorama, 5062).
- Gaude, J. Evaluating public training and employment programmes. In: *Evaluation of European training, employment and human resource programmes*. Luxembourg: Office for Official Publications of the European Communities, 1997, p. 51-58 (Cedefop Panorama, 5062).
- Gontzou, C. Methodological aspects of European training and employment programmes. In: *Evaluation of European training, employment and human resource programmes*. Luxembourg: Office for Official Publications of the European Communities, 1997, p. 61-62 (Cedefop Panorama, 5062).
- Grubb, N. W.; Ryan, P. *The roles of evaluation for vocational education and training*. Geneva: International Labour Office, 1999.
- Guba, Y.; Lincoln, E. *Forth generation evaluation*. London: Sage, 1989.
- Gullickson A R. *The student evaluation standards: how to improve evaluations of students*. Thousand Oaks: Corwin Press, 2002.
- Hägele, H. Experteninterviews in der öffentlichen Verwaltung: ausgewählte praktische Probleme. In: Brinkmann, Ch. et al. (eds) *Experteninterviews in der Arbeitsmarktforschung, Beiträge zur Arbeitsmarkt- und Berufsforschung*. 1995, No 191, p. 69-72.
- Hale, J. *Performance-based evaluation: tools and techniques to measure the impact of training*. San Francisco: Jossey-Bass 2002.

- Haller, S. *Beurteilung von Dienstleistungsqualität. Dynamische Betrachtung des Qualitätsurteils im Weiterbildungsbereich*. Wiesbaden: 1998.
- Heckman, J. J. Micro data, heterogeneity and the evaluation of public policy: Nobel Lecture. *Journal of Political Economy*, 2001, Vol. 109, No 4, p. 673-748.
- Heckman, J. J.; Smith, J. A. Experimental and nonexperimental evaluation. In: Schmid, G. et al. (eds) *International handbook of labour market policy and evaluation*. Cheltenham: Edward Elgar, 1996, p. 37-87.
- Hendricks, M.; Conner, R. F. International perspectives on the guiding principles. Shadish, W. R. et al. (eds) *Guiding principles for evaluators. New directions for program evaluation*. San Francisco: Jossey-Bass, Summer 1995, No 66, p. 77-90.
- Hofstede, G. *Culture's consequences: international differences in work-related values*. London et al.: Sage Publications, 1980.
- House, E. R.; Howe, K. R. Deliberative democratic evaluation. *New Directions for Evaluation*, 2000, No 85, p. 3-12.
- Hujer, R. et al. Evaluation aktiver Arbeitsmarktpolitik – Probleme und Perspektiven. *MittAB*, 2000, No 3, p. 341-344.
- ILO – International Labour Office. *Guidelines for the preparation of independent evaluations of ILO programmes and projects*. Last update 15 November 1999. Available from Internet: <http://www.ilo.org/public/english/bureau/program/guides/indpen/index.htm> [cited 3.11.2003].
- Jacob, St.; Varone, F. *L'évaluation au concret en Belgique: méta-évaluation au niveau fédéral*. Deuxième rapport intermédiaire du projet de recherche AM/10/016 par les SSTC, provisional version, May 2002.
- James, C.; Roffe, I. The evaluation of goal and goal-free training innovation. *Journal of European Industrial Training*, 2000, Vol. 24, No 1, p. 12-20.
- Jang, S. The appropriateness of Joint Committee standards in non-western settings: a case study of South Korea. In: Russon, C. (ed.) *The program evaluation standards in international settings*. Kalamazoo, MI: The Evaluation Center, 2000, p. 41-59 (Occasional Papers Series, May 1).
- Jesse, E. Typologie politischer System der Gegenwart. In: Bundeszentrale für politische Bildung (ed.) *Grundwissen Politik*. Bonn: 1993, p. 165-227.
- JC – Joint Committee on Standards for Educational Evaluation (ed.) *The program evaluation standards. How to assess evaluations of educational programs*. Thousand Oaks: Sage, 1994.
- JC – Joint Committee on Standards for Educational Evaluation (ed.) *Standards for evaluation of educational programs, projects and materials*. New York: JC, 1981.
- JC – Joint Committee on Standards for Educational Evaluation (ed.) *The personnel evaluation standards*. Newbury Park, CA: Sage Publications, 1988.
- JC – Joint Committee on Standards for Educational Evaluation (ed.) *Handbuch der Evaluationsstandards*. Opladen: Leske und Budrich, 2000.
- JC – Joint Committee on Standards for Educational Evaluation
- Kaiser, F.-J. Fremdevaluation: Inwieweit sind die Erkenntnisse aus Modellversuchen inhaltlich und methodologisch für die Berufsbildungsforschung verwendbar? In: Euler, D. (ed.) *Berufliches Lernen im Wandel – Konsequenzen für die Lernorte?* Beiträge zur Arbeitsmarkt- und Berufsforschung, 1998, No 214, p. 537-550.
- Kellaghan, T.; Stufflebeam, D. L. (eds) *International handbook of educational evaluation*. Dordrecht: Kluwer, 2002.
- Kirkpatrick, D. L. *Evaluating training programs: the four levels*. San Francisco, CA: Berrett Koehler, 1994.
- Klieme, E. Bildungsstandards als Beitrag zur Qualitätsentwicklung im Schulsystem. *DIPF informiert*, August 2002, No 3, p. 2-6.
- Kluge, F. *Etymologisches Wörterbuch der deutschen Sprache*. Berlin/New York: de Gruyter 1999.
- Knox, A. B. Evaluation of continuing education in the USA. In: Künzel, K. (ed.) *Evaluation der Weiterbildung*. Cologne: Böhlau, p. 201-213 (Internationales Jahrbuch der Erwachsenenbildung, Vol. 27)
- Kuffner, A. *Evaluation von Nachhaltigkeitsaspekten – Nachhaltige Evaluation?* Dissertation University of Vienna, 2000 (unpublished).
- Kushner, S. *Personalizing evaluation*. Sage: Thousand Oaks, 2000.
- Leeuw, F. L. Evaluation in Europe. In: Stockmann, R. (ed.) *Evaluationsforschung, Sozialwissenschaftliche Evaluationsforschung 1*. Opladen: Leske und Budrich, 2000, p. 57-76.

- Levin, H. M.; McEwan, P. J. *Cost-effectiveness analysis*. London: Sage, 2001.
- Lindley, R. M. The European Social Fund: a strategy for generic evaluation In: Schmid, G., O'Reilly, J.; Schömann, K. (eds) *International handbook of labour market policy and evaluation*. Cheltenham: Edward Elgar, 1996. p. 843-867.
- Luschei, F.; Trube, A. Evaluation und Qualitätsmanagement in der Arbeitsmarktpolitik – Einige systematische Vorüberlegungen und praktische Ansätze zur lokalen Umsetzung. *MittAB*, 2000, No 3, p. 533-549.
- Madaus, G. F.; Stufflebeam, D. L. *Educational evaluation: the classical writings of Ralph W. Tyler*. Boston: Kluwer, 1988.
- Mark, M. M.; Henry, G. T.; Julnes, G. *Evaluation: an integrated framework for understanding, guiding, and improving policies and programs*. San Francisco: Jossey-Bass, 2000.
- Marklund, S. Applicability of standards for evaluations of educational programs, projects and materials in an international setting. *Evaluation and Program Planning*, 1984, Vol. 7, p. 355-362.
- Nuissl, E. *Adult education and learning in Europe – Evaluation of the adult education action within the Socrates programme*. Frankfurt a.M.: DIE – Deutsches Institut für Erwachsenenbildung, 1999.
- Oliva, D.; Samek Lodovici, M. Le politiche formative tra occupazione e valorizzazione delle risorse umane. *Rassegna Italiana di Valutazione*, 1999, No 13. Available from Internet: <http://www.valutazioneitaliana.it/riv/rivista99/13-olivasamek.doc> [cited 6.11.2003].
- Owen, J. M.; Rogers, P. J. *Program evaluation*. London: Sage Publications, 1999.
- Patton, M. Q. *Utilization-focused evaluation: the new century text*. Thousand Oaks: Sage 1997.
- Pawson, R.; Tilley, N. *Realistic Evaluation*. Thousand Oaks: Sage, 1997.
- Peltzer, U. Formative Prozessbegleitung des Forschungs- und Entwicklungsprogramms 'Lernkultur Kompetenzentwicklung'. *Berufliche Kompetenzentwicklung Bulletin*, 2002, No 2, p. 8-10.
- Perret, B.; Barbier, J.-C. [2000 update]. *Ethical Guidelines, Process and Product Quality Standards, What For?* An SFE (French Evaluation Society) Perspective. Paper presented at the European Evaluation Society Conference in Lausanne, 12-14 October 2000. Available from Internet: http://www.europeanevaluation.org/pdf/6-3_barbier-perret.pdf [cited 4.11.2003].
- Polverari, L.; Fitzgerald, R. *Integrating Gender Equality in the Evaluation of the Irish 2000-06 National Development Plan, Vol. 2: Tool Kit for Gender Evaluation*. Glasgow: 2002.
- Raabe, B. *Wirkungen aktiver Arbeitsmarktpolitik. Evaluierungsergebnisse für Deutschland, Schweden, Dänemark und die Niederlande*. Wissenschaftszentrum Berlin für Sozialforschung, July 2000 (Discussion Paper FS I 00-208).
- Reischmann, J. *Weiterbildungs-Evaluation*. Neuwied, Kriftel: Luchterhand, 2003.
- Rieper, O. Evaluation practice of European Structural Funds. In: *Evaluation of European training, employment and human resource programmes*. Luxembourg: Office for Official Publications of the European Communities, 1997, p. 37-43 (Cedefop Panorama, 5062).
- Rossi, P. H.; Freeman, H. E.; Lipsey, M. W. *Evaluation a systematic approach*. Thousand Oaks: Sage, 1999.
- Rost, J. Allgemeine Standards für die Evaluationsforschung. Hager, W. et al. (eds) *Evaluation psychologischer Interventionsmaßnahmen*. Bern: Hans Huber, 2000, p. 129-140.
- Russon, C; Russon, K. (eds) *The annotated bibliography of international programme evaluation*. Dordrecht: Kluwer, 2000.
- Sanders, J. R. Standards and Principles. In: Shadish, W. R. et al. (eds) *Guiding Principles for Evaluators, New Directions for Program Evaluation*. San Francisco: Jossey-Bass, Summer 1995, No 66, p. 47-52.
- Sauter, E.; Schmidt, H. *Training standards in Germany. The development of new vocational education and training standards*. Bonn: Federal Institute for Vocational Training, 2002.
- Schmid, G. Process evaluation: policy formation and implementation. In: Schmid, G. et al. (eds) *International handbook of labour market policy and evaluation*. Cheltenham: Edward Elgar, 1996, p. 198-231.
- Schmid, G. et al. (eds) *International handbook of labour market policy and evaluation*. Cheltenham: Edward Elgar, 1996.
- Schmidt, Chr. Arbeitsmarktpolitische Maßnahmen und ihre Evaluierung: Eine Bestandsaufnahme. *DIW Vierteljahresheft*, 2000, No 3, p. 425-437.

- Schmidt, Ch. *Knowing what works: the case for rigorous program evaluation*. London: CEPR, 2001 (CEPR Working paper).
- Schmitt von Sydow, H. *Report of the working group 'Evaluation and transparency*. White Paper on European governance, Work Area No 2: Handling the process of producing and implementing community rules. July 2001. Available from Internet: http://europa.eu.int/comm/governance/areas/group4/report_en.pdf [cited 3.11.2003].
- Schömann, K. Longitudinal designs in evaluation studies. In: Schmid, G. et al. (eds) *International handbook of labour market policy and evaluation*. Cheltenham: Edward Elgar, 1996, p. 115-142.
- Sen, A. K. *The standard of living*. Cambridge 1987.
- Seyfried, E. *Evaluation of quality aspects in vocational training programmes*. Luxembourg: Office for Official Publications of the European Communities, 1998. (Cedefop Document, 1171).
- Shadish, W. R. et al. (eds) *Guiding principles for evaluators, New Directions for Program Evaluation*. San Francisco: Jossey-Bass, Summer 1995, No 66.
- Shadish, W. R.; Cook, T. D.; Campbell, D. T. *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin, 2002.
- Sloane, P. F. E. Das Potential von Modellversuchsfeldern für die wissenschaftliche Erkenntnisgewinnung. In: Bentler, P. et al. (eds) *Modellversuchsforschung als Berufsbildungsforschung*. Cologne: Böhlau, 1995, p. 11-44.
- Smith, N. L. et al. Considerations on the development of culturally relevant evaluation standards. *Studies in Educational Evaluation*, 1993, Vol. 19, No 1, p. 3-13.
- Smith, N. L.; Jang, S. Increasing cultural sensitivity in evaluation practice. *Studies in Educational Evaluation*, 2002, Vol. 28, No 1, p. 61-69.
- SFE – Société Française de l'Évaluation. *Projet de 'texte préliminaire pour une charte'*, présenté au colloque SFE, June 2001 (internal document, unpublished).
- Speer, S. Evaluation und Benchmarking – Ein Methodenvergleich am Beispiel der betrieblichen Personalarbeit. In: Geise, W. (ed.) *Ökonomische Bildung zur Bewältigung von Lebenssituationen*. Bergisch-Gladbach: Hobein, 2001, p. 51-67.
- Stake, R. E. *The art of case study research*. Sage: Thousand Oaks, 1995.
- Stockdill, S.H. *Evaluation standards. A study of their appropriateness in business and education*. University of Minnesota: DAI – Dissertation Abstracts International, 1986 (47-08 A:2967).
- Stufflebeam, D. L. *Standards of practice for evaluators*. Lecture at the annual conference of the American Educational Research Association, San Francisco, 1986.
- Stufflebeam, D. L. Evaluation Models. In: *New Directions for Evaluation*. San Francisco: Jossey-Bass, Spring 2001, No 89.
- Stufflebeam, D. L. et al. *Educational evaluation and decision making*. Itasca, Ill.: Peacock, 1971.
- Taut, S. Cross-cultural transferability of the program evaluation standards. In: Russon, C. (ed.) *The program evaluation standards in international settings*. Kalamazoo, MI: The Evaluation Center, 2000, p. 5-27 (Occasional Papers Series, May 1).
- Tenberg, R. Selbstevaluation des Modellversuchs Fächerübergreifender Unterricht in der Berufsschule durch den Lehrstuhl für Pädagogik der Technischen Universität München. In: Euler, D. (ed.) *Berufliches Lernen im Wandel – Konsequenzen für die Lernorte?* Beiträge zur Arbeitsmarkt- und Berufsforschung, 1998, No 214, p. 527-535.
- Toulemonde, J. Evaluation culture(s) in Europe: differences and convergence between national practices. *DIW Vierteljahresheft*, 2000, No 3, p. 350-357.
- Tremea – Training Effectiveness Measurement. *Tremea Handbook. A Guide for Evaluating Training Programmes*. 2002. Available from Internet: www.tremea.gr [cited 3.11.2003]
- Turbin, J. *Policy borrowing: lessons from European attempts to transfer training practices*. Leicester: Centre for Labour Market Studies, 2000 (CLMS Working Paper No 27)
- UKES – United Kingdom Evaluation Society. *Guidelines for good practice in evaluation*. Available from Internet: http://www.evaluation.org.uk/ukes_new/Pub_library/Guidance-Good Practice.doc [cited October 2002].
- Uusikyla, P; Virtanen, P. Meta-evaluation as a tool for learning: a case study of the European structural fund evaluations in Finland. *Evaluation*, 2000, Vol. 6, Issue 1, p. 50-65.

- Vedung, E. *Evaluation im öffentlichen Sektor*. Vienna: Böhlau, 1999.
- Villa, A. *Estándares para la Evaluación de Programas*. Bilbao: Ediciones Mensajero, 1998.
- Villa, A. *Estándares de Evaluación de Personal*. Bilbao: Ediciones Mensajero, 1999.
- Webster's encyclopedic unabridged dictionary of the English language. New York: Thunder Bay Press, 1989.
- Weiß, R. Methoden und Faktoren der Erfolgsmessung in der betrieblichen Weiterbildung. *GdWZ*, 1997, Vol. 3, p. 104-108.
- Widmer, Th. *Meta-Evaluation: Kriterien zur Bewertung von Evaluationen*. Bern: Haupt, 1996.
- Widmer, Th. Kontext, Inhalt und Funktion der 'Standards' für die Evaluation von Programmen. In: Müller-Kohlenberg, H.; Münstermann, K. (eds) *Qualität von Humandienstleistungen*. Opladen: Leske und Budrich, 2000, p. 77-88.
- Widmer, Th. Instruments and procedures for assuring evaluation quality: a Swiss perspective. In: Schwartz, R. Mayne, J. Toulemonde, J. (eds) *Assuring the quality of evaluative information: prospects and pitfalls*. New Brunswick: Transaction, 2003 (forthcoming).
- Widmer, Th.; Beywl, W. Die Übertragbarkeit der Evaluationsstandards auf unterschiedliche Anwendungsfelder. In: JC; Sanders, J. R. (eds) *Handbuch der Evaluationsstandards*, Opladen: Leske und Budrich, 2000, p. 243-257.
- Widmer, T.; Landert, Ch.; Bachmann, N. *Evaluations standards*. Genève: SEVAL – Swiss Evaluation Society, December 2000. Available from Internet: http://www.seval.ch/en/documents/SEVAL_Standards_2000_en.pdf [cited 4.11.2003].
- Wingens, M.; Sackmann, R. Evaluation AFG-finanzierter Weiterbildung. *MittAB*, 2000, No 1, p. 39-53.
- Wollmann, H. Policy knowledge and contractual research. *International Encyclopedia of Social and Behavioral Sciences*, 2002, Vol. 5.4 (forthcoming).
- Wottawa, H. Evaluation in der betrieblichen Bildung. In: Künzel, K. (ed.) *Evaluation der Weiterbildung*. Cologne: Böhlau, 1999, p. 105-116. (Internationales Jahrbuch der Erwachsenenbildung, Vol. 27)
- Wottawa, H.; Thierau, H. *Lehrbuch evaluation*. Bern: Hans Huber, 1998.
- Zimmer, G. Durch Modellversuche zu Erkenntnisgewinn und Praxisinnovation? Zur Positions-, Funktions- und Interessenbestimmung der wissenschaftlichen Begleitforschung. In: Euler, D. (ed.) *Berufliches Lernen im Wandel – Konsequenzen für die Lernorte?* Beiträge zur Arbeitsmarkt- und Berufsforschung, 1998, No 214, p. 596-607.
- Zorzi, R. et al. Canadian evaluation society project in support of advocacy and professional development. Evaluation benefits, outputs, and knowledge elements. Toronto: Zorzi and Associates, October 2002. Available from Internet: <http://consultation.evaluationcanada.ca/pdf/ZorziCESReportDuplex.pdf> [cited 04.11.2003].

Methods and limitations of evaluation and impact research

Reinhard Hujer, Marco Caliendo,
Dubravko Radić ⁽¹⁾

Abstract

The need to measure and judge the effects of social programmes, and the importance of evaluation studies in this context, is no longer questioned. Evaluation is a complex task and involves several steps. The following contribution mainly focuses on the methodological aspects of evaluation, particularly on econometric evaluation techniques. In consequence, we concentrate on evaluation studies conducted in the field of active labour-market policies (ALMP) and especially labour-market training (LMT) in Germany and Europe. The ideal evaluation process can be viewed as a series of three steps. First, the impacts of the programme on the individual should be estimated. Second, it should be determined whether the estimated impacts are large enough to yield net social gains. Finally, evaluation should question whether the best outcome have been achieved for the money spent (Fay, 1996). The focus of our paper is the first step, namely the microeconomic evaluation, although the other two steps, i.e. macroeconomic evaluation and cost-benefit analysis (CBA), are also discussed.

When discussing microeconomic evaluation, analysts such as LaLonde (1986) or Ashenfelter and Card (1985) view social experiments as the only valid evaluation method. A second group, including Heckman and Hotz (1989) and Lechner (1998), believe that it is possible to construct a comparison group using non-experimental data and econometric and statistical methods to solve the fundamental evaluation problem. This problem arises because we cannot observe individuals simultaneously with and without participation in a programme.

Since experimental data is scarce in Europe and evaluators often have to evaluate programmes already running, we focus on the techniques that deal with non-experimental data only. To do so we discuss the fundamental evaluation problem and different methods to solve this problem empirically. The methods presented include the before-after estimator (BAE), the cross-section estimator (CSE), the matching estimator, the difference-in-differences estimator (DID) and finally the duration model approach. Every estimator makes some generally untestable assumption to overcome the fundamental evaluation problem and therefore we also discuss the likelihood that these assumptions are met in practice. In addition we present a numerical example to show how the estimators are implemented technically and also discuss their data requirements. Aggregated data can be used as well as individual data to evaluate the effects of ALMP programmes (including LMT). Instead of looking at the effect on individual performance, we would like to know if the ALMP represent a net gain to the whole economy. If the total number of jobs is not affected by labour-market policies, the effects will only be distributional. The need for a macroeconomic evaluation arises from the presence of dead-weight losses, displacement and substitution effects. Important

⁽¹⁾ Acknowledgements: the authors thank Björn Christensen, Pascaline Descy, Viktor Steiner, Manfred Tessaring, Stephan Thomsen and Christopher Zeiss for valuable comments and Yasemin Illgin, Paulo Rodrigues and Oliver Wünsche for research assistance.

methodological issues in a macroeconometric evaluation are the specification of the empirical model, which should always be based on an appropriate theoretical framework, and the simultaneity problem of ALMP which has to be solved.

Despite being able to reveal the impacts for the individuals and for the whole economy, micro- and macroeconometric evaluations do not cover the full range of effects associated with social programmes. Social programmes also involve various other, more qualitative, aims and objectives. Examples include distributional and equity goals, which are often hard to measure but nevertheless important. In order to capture these effects, a cost-benefit analysis should be conducted as a third step. Such an analysis widens the perspective of an impact analysis and contributes to better understanding of the complete effects of a social programme.

After having discussed the methodological issues related to the three evaluation steps, we present empirical findings from micro- and macroeconometric evaluations in Europe and, in particular, in Germany. We find that training programmes seem to have positive effects on the individuals in most of the studies and perform better than alternative labour-market programmes such as, for example, job creation schemes (JCS).

Finally, we use the results from our methodological discussion and the empirical findings to draw some implications for policy and evaluation practice. We focus on the choice of the appropriate estimation method, data requirements, the problem of heterogeneity in evaluation analysis, the successful design of training programmes and the transferability of our findings (for ALMP) to other social programmes.

Table of contents

1. Introduction	136
2. Methods of evaluation	139
2.1. Process of evaluation	139
2.1.1. The evaluand and the time of the evaluation	139
2.1.2. The evaluators	140
2.1.3. The evaluation criteria	140
2.1.4. The evaluation method	141
2.2. Microeconomic evaluation	142
2.2.1. Fundamental evaluation problem	142
2.2.2. Before-after estimator	144
2.2.3. Cross-section estimator	146
2.2.4. Matching estimator	146
2.2.5. Difference-in-differences estimator	149
2.2.6. Duration models	150
2.2.7. An intuitive example	151
2.3. Macroeconomic evaluation	154
2.4. Cost-benefit analysis (CBA)	156
2.4.1. Time of a CBA	158
2.4.2. Selection of the alternatives	158
2.4.3. Defining the reference group	158
2.4.4. Enumerate and forecast all impacts	159
2.4.5. Measuring and aggregating impacts	160
2.4.6. Comparing costs and benefits	161
2.4.7. Conducting sensitivity tests	161
3. ALMP in the EU and empirical findings	165
3.1. Recent developments in ALMP in the EU	165
3.2. Microeconomic evaluation studies	167
3.2.1. Germany	167
3.2.2. Other European countries	168
3.2.2.1. Switzerland	168
3.2.2.2. Austria	169
3.2.2.3. Belgium	169
3.2.2.4. France	169
3.2.2.5. Denmark	169
3.2.2.6. United Kingdom	169
3.2.2.7. Ireland	169
3.2.2.8. Norway	169
3.2.2.9. Sweden	170
3.3. Macroeconomic evaluation studies	170
3.3.1. Sweden	170
3.3.2. Germany	170
3.4. Vocational training programme success in previous years	171

4.	Policy implications: some guidance for evaluation and implementation practice	175
4.1.	Selection problem and the choice of estimation method	175
4.2.	Heterogeneity	176
4.3.	Data requirements	176
4.4.	Macroeconometric analysis and cost-benefit analysis	177
4.5.	The design of training programmes: some suggestions	178
4.6.	Transferability to other social programmes	179
5.	Summary and conclusions	180
	List of abbreviations	182
	Annex: Tables	183
	References	186

List of tables and figures

Tables

Table 1: Comparison of the different evaluation estimators	151
Table 2: Labour-market outcomes for participants and non-participants	152
Table 3: Estimates of the programme effects	154
Table 4: Hypothetical example before reallocation	162
Table 5: Hypothetical example after reallocation	162
Table 6: Costs and benefits of the Job Corps programme	163
Table 7: Expenditure on LMT as a percentage of total spending on ALMP	167
Table 8: Summary of the empirical findings from microeconomic studies	172
Table A1: Standardised unemployment rates in the EU (as a percentage of total labour force)	183
Table A2: Public expenditure on ALMP as a percentage of GDP	184
Table A3: Public expenditure on PLMP as a percentage of GDP	184
Table A4: Public expenditure on LMT as a percentage of GDP	185

Figures

Figure 1: Spending on ALMP, PLMP and LMT (EU average) and unemployment rates (EU average) from 1985 to 2000	137
Figure 2: Major aspects of an evaluation process	141
Figure 3: Ashenfelter's dip	145
Figure 4: Different stages of a CBA	157
Figure 5: Spending on ALMP and PLMP in the EU, 1985	166
Figure 6: Spending on ALMP and PLMP in the EU, 2000	166

1. Introduction

The need to measure and judge the effects of social programmes, and the importance of evaluation studies in this context, is no longer questioned. Evaluation is, however, a complex task and involves several steps. The range of topics is virtually unlimited and, since every topic requires a specific methodological approach, there exists no general evaluation strategy. To be successful, every evaluation must pre-specify a set of preliminary aspects. First it must be stated precisely which subject is to be evaluated. A second important question is 'who are the evaluators and what are their competences?'. Finally, and probably most important, the evaluation criteria and the chosen procedure for the evaluation must be fixed.

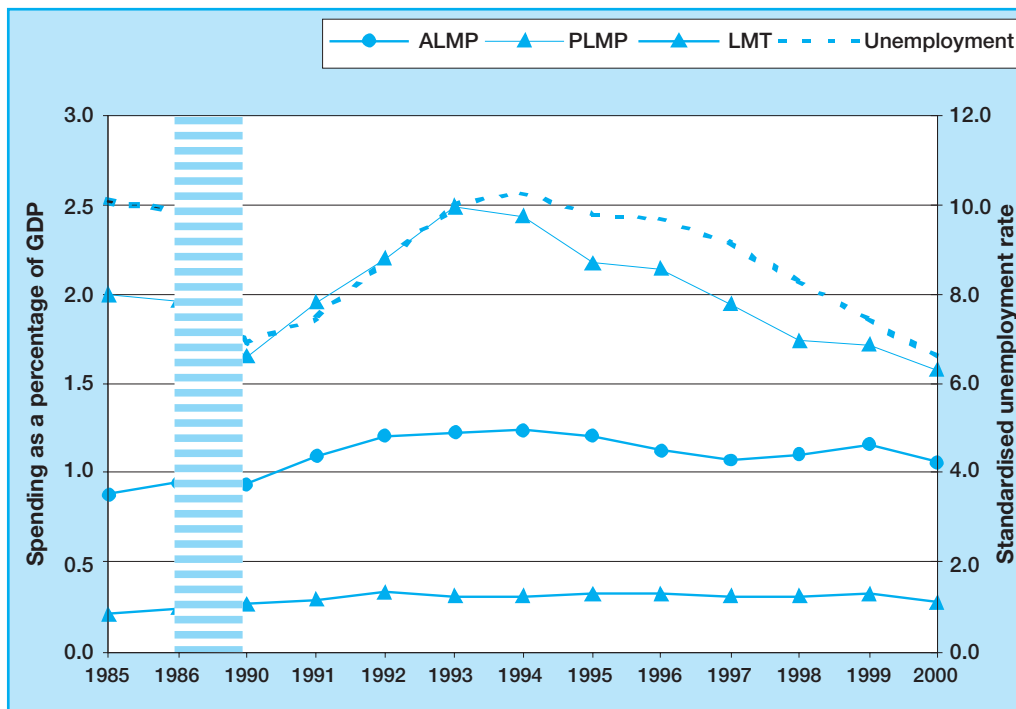
This contribution will focus mainly on the last aspect, i.e. offer some suggestions and advice on which econometric evaluation techniques should be used under different economic circumstances. In doing so, we will concentrate on evaluation studies conducted in the field of active labour-market policies (ALMP) and especially labour-market training (LMT). We regard them as fruitful object of analysis for several reasons. The increasing expenditure on ALMP in the last decades has raised a considerable interest in evaluation and motivated many studies in that field. Also the choice of the outcome variable that is of interest is more obvious than for other social programmes; the question of what should be defined as a success and how to measure it can be answered more easily. Whereas evaluation in the United States, for example, has mainly focused on earnings after participating in a certain programme, in Europe the focus is more on the employment prospects of the participants. We will address the issue of choosing an appropriate outcome variable later on. Finally, in contrast to health programmes, for example, there is a closer link between ALMP programmes, which are often short-term activities, and the outcome considered. Therefore, the evaluation of the effect of these programmes is often easier.

ALMP have been seen as one way to fight unemployment, rising in most European countries since the early 1970s. The growing interest in these measures is easy to understand in view of the disillusionment with more aggregate policies. Traditional demand stimulation has been discredited because it faces the risk of increasing inflation with only small effects on employment. Supply-side structural reforms, aimed at removing various labour-market rigidities, are difficult to implement or appear to produce results rather slowly. In this situation, as Calmfors (1994) notes, ALMP are regarded by many as the *deus ex machina* that will provide the solution to unemployment. Not only do they provide a more efficient outcome in the labour market, they also equip individuals with higher skills and therefore lower the risk of poverty. In this sense ALMP are capable of meeting efficiency and equity goals at the same time (OECD, 1991) ⁽²⁾. Another development that spurred interest in ALMP is the fact that it has become a common theme in the political debate that governments should shift the balance of public spending on labour-market policies away from passive income support towards more active measures designed to get the unemployed back into work. Anglo-Saxon policy-makers especially favour the idea of tying the right of welfare to the duty of work. Welfare then becomes workfare (Card, 2000).

This should manifest itself in a higher relative importance of ALMP. Figure 1 shows the average spending on active (including LMT) and passive labour-market policies (PLMP) as a percentage of the national gross domestic product (GDP) from 1985 to 2000 for the countries of the European Union (EU). The average spending on ALMP rose from 0.88 % of GDP in 1985 to 1.23 % in 1994. After that peak it went back to 1.06 % in 2000. The spending on PLMP, in contrast, peaked in 1993 at 2.49 % before it was reduced to 1.57 % in 2000. If one compares the relationship between the two measures in 1985 and 2000 the growing impor-

⁽²⁾ Several means by which ALMP might influence the labour markets, such as enhancing and adapting skills according to skill needs, improving individual employability and avoiding skill shortages, can be thought of.

Figure 1: Spending on ALMP, PLMP and LMT (EU average) and unemployment rates (EU average) from 1985 to 2000



tance of active measures becomes clearer. Whereas in 1985 the spending on PLMP was 126 % higher than the spending on ALMP, it was only 48 % higher in 2000. Even though this is a remarkable change in the structure of the spending, high amounts are still directed to passive measures.

One obvious reason for the limited success in switching resources into active measures is the fact that unemployment benefits are entitlement programmes. Rising unemployment automatically increases public spending on passive income support, whereas most of the active labour-market programmes are discretionary in nature and therefore easier to dispense with in a situation of tight budgets (Martin, 1998). This becomes clear from examination of Figure 1 which also shows standardised unemployment rates from 1985 to 2000. The EU average peaked at 10.1 % in 1985, 6.9 %

in 1991 and 10.3 % in 1994. After that unemployment went down to 6.9 %. The movement of the standardised unemployment rates and spending on PLMP is nearly synchronous. LMT plays an important role as one measure of ALMP. The average spending on LMT in the EU rose from 0.21 % in 1985 to 0.28 % in 2000.

Clearly, the money spent on labour-market programmes is not available for other projects or private consumption. In an era of tight government budgets and a growing disbelief regarding the positive effects of ALMP ⁽³⁾, evaluation of these policies becomes imperative. Various evaluation methods have been developed for this. The ideal evaluation process can be viewed as a series of three steps. First, the impacts of the programme on the individual or groups of individuals should be estimated. Second, it should be

⁽³⁾ There have been several microeconomic studies which showed no positive effects of ALMP, especially for job creation schemes. See Hujer and Caliendo (2001) for an overview of the evaluation studies on vocational training and job creation schemes in West and East Germany.

determined whether the estimated impacts are large enough to yield net social gains. Finally, it should be decided if this is the best outcome that could have been achieved for the money spent (Fay, 1996) ⁽⁴⁾. The focus of our paper is the first step, namely the microeconomic evaluation, although the other two steps, i.e. macroeconomic evaluation and cost-benefit analysis (CBA), are discussed too.

Empirical microeconomic evaluation is conducted with individual data. The main question is whether the desired outcome variable for an individual is affected by participation in an ALMP programme. Relevant outcome variables could be the future employment probability or the future earnings. We would like to know the difference between the value of the participants' outcome in the actual situation and the value of the outcome if there had been no participation in the programme. The fundamental evaluation problem arises because we never observe both states (participation and non-participation) for the same individual at the same time. Therefore, finding an adequate control group is necessary to make a comparison possible. This is not easy because participants in programmes usually differ systematically from the non-participants in more aspects than just participation. Simply taking the difference between their outcomes after the programme will not reveal the true programme impact but will lead to a biased estimate. The literature on the solution to this problem is dominated by two points of view.

Analysts such as LaLonde (1986) or Ashenfelter and Card (1985) view social experiments as the only valid evaluation method. A second group, including Heckman and Hotz (1989) and Lechner (1998), believe that it is possible to construct a comparison group using non-experimental data and econometric and statistical methods to solve the fundamental evaluation problems. In non-experimental or observational studies, the data are not derived in a

process that is completely under the control of the researcher. Instead, one has to rely on information about how individuals actually performed after the intervention; that is, we observe the outcome with treatment for participants and the outcome without treatment for non-participants.

The objective of observational studies is to use this information to restore the comparability of both groups by design. To do so, more-or-less plausible identification assumptions have to be imposed. There are several approaches differing with respect to the methods applied to this problem. Some studies control for observables as part of parametric evaluation models, others construct matched samples. Some authors think that conditioning on observables is not enough and that one has to take into account unobservables too. We present here several different estimation techniques, discuss the methodological concepts associated with them, highlight their (dis-) advantages and identify the environments under which they work best. While presenting the methodological concepts of each estimator, we also discuss the data requirements needed for their implementation. By doing this, we hope to give advice for the construction of datasets in the future.

The remainder of this paper is organised as follows:

- (a) first, we present some methodological concepts of evaluation. We discuss microeconomic evaluation concepts in detail and also present some ideas on macroeconomic analysis and the cost-benefit approach;
- (b) in the following chapter we present some standardised facts about the evolution of unemployment and labour-market policies in the EU, before we give an overview of the most relevant previous empirical findings regarding LMT in Germany and Europe;
- (c) finally, we summarise the findings and give advice for the implementation and evaluation of training or other social programmes in the future.

⁽⁴⁾ The results can also be used to provide a feedback for improvement of subsequent programmes.

2. Methods of evaluation

2.1. Process of evaluation

The need to measure and judge the effects of social programmes, and the importance of evaluation studies in this context, is no longer questioned. Founders of these programmes increasingly ask for hard evidence of the efficient use of investments and enquire whether the programmes were successful or not. Evaluation is, however, a complex task that involves several steps. The range of topics is virtually unlimited and, since every topic requires a specific methodological approach, there is no general evaluation strategy.

Evaluation is, according to a definition of Worthen et al. (1997), the determination of the worth or merit of an evaluation object, the *evaluand*. It comprises (Worthen et al., 1997, p. 5) the 'identification, clarification, and application of defensible criteria to determine an evaluation object's value, quality, utility, effectiveness, or significance in relation to those criteria.' Its outstanding attribute, therefore, is its scientific claim. Another feature of evaluation that distinguishes it from basic research is its practical access. Evaluation and research are similar in that they use empirical methods and techniques to discover new knowledge. However, whereas research is primarily interested in advancing knowledge for its own sake, evaluation of social programmes is concerned with practical utilisation and matters such as:

- (a) did the programme achieve what it was intended to do?
- (b) who benefited from the programme?
- (c) could the programme be conducted better or more efficiently?
- (d) what changes must be undertaken to improve particular aspects of the programme?

To be successful, every evaluation must pre-specify a set of preliminary aspects (Kromrey, 2001). First, it must be stated precisely which subject should be evaluated and when this evaluation should take place. A second important point is to consider who the evaluators are and what their competences are. Finally, and probably most

important, the evaluation criteria and the chosen procedure for the evaluation must be fixed.

2.1.1. The *evaluand* and the time of the evaluation

Defining the evaluation object seems straightforward at first sight. It comprises a detailed description of the programme which should be evaluated. A broad variety of programmes exists: those already implemented; those at the implementation stage; regionally concentrated pilot-programmes; and well established national programmes. However, a detailed description of the programme is not enough. Since an evaluation of every aspect of the programme would not be feasible, it must be clearly specified whether only the implementation of the programme should be evaluated or certain impacts of it.

Monitoring the implementation of the programme consists of a complete enumeration of what has happened during the different stages of its execution. It gives a first impression of the programme under consideration, for example the number of participants, the average time the agencies conducting the programmes spent with each client, the expected costs of the programme, the completion rate, the employment status reached after the participation, etc. Such information can be a form of control over the agents implementing the programme. Although this monitoring gives a first impression about the success or failure of a programme, it does not give any explanations for it. Monitoring is thus a reasonable first step in an evaluation process and can be seen as a minimum requirement to check how large sums of public money are spent. Evaluation, on the other hand, goes a step further and aims at determining whether a programme is successful or not by defining certain criteria and assessing whether these criteria were met.

Another question which has to be answered before the evaluation takes place concerns the time when an object should be evaluated. In this context, a distinction between formative and summative evaluation seems to be useful. If the evaluation takes place during the development

stage of the programme, and if the results of the evaluation are used to provide some feedback on how to improve it, a formative evaluation is conducted. Such an open formative evaluation is of practical value because of its feedback feature. However, the feedback of evaluation results on the programme does not allow for interpretation of the results in terms of success or efficiency since the evaluation itself influences the success of the programme. In addition, most programmes yield benefits only in the medium to long term and hence cannot be used as feedback for current programmes but, at best, as a basis for future improvements of similar programmes.

A summative evaluation, on the other hand, is conducted after the programme has been finished. Thereby an immediate influence on the results on the programme is abandoned. The major task of a summative evaluation is to determine whether the programme should be continued or terminated after it was carried out. Scriven (1991) puts the difference between the two approaches using the following illustration: 'When the cook tastes the soup, that's formative evaluation; when the guest tastes it, that's summative evaluation.'

2.1.2. The evaluators

Another important point is the question of who is authorised to conduct an evaluation. In an internal evaluation, the employees working for the programme are in charge of it. The alternative would be to assign this task to external evaluators, for example independent consultants or researchers. The advantage of an internal process is that the evaluators are familiar to the programme and that they have trouble-free access to all necessary information. One potential problem is the desiderative professionalism and objectivity. If, for example, the institution in charge of the evaluation is also responsible for the implementation of the programme, an incentive to find results that correspond to the aims and objectives of the programme could arise. This hazard can be avoided by external evaluators who are also able to bring in new views and ideas. But, even if the evaluation is done by internal staff, the transparency and accountability can be increased if there is some kind of cooperation with external institutions in certain areas, or if the material underlying the evaluation, for example datasets, etc., are made available to the scientific community.

2.1.3. The evaluation criteria

The *a-priori* fixing of criteria ensures that the evaluation is not done in an ad hoc manner but transparently and comprehensibly. There is a broad spectrum of potential criteria, including various direct and secondary impacts of the programme. In the case of ALMP for example, direct impacts could be on future employment probability or wages, whereas secondary effects would include potential displacement and crowding-out-effects. We will address these issues shortly in more detail. Other criteria could be the efficiency in conducting the programme or even the legitimacy of the objectives themselves. In this context an important question is where these criteria come from, i.e. who sets them?

Typically, the criteria stem from the programme itself. If for example the programme to be evaluated aims at improving the re-employment probability of the disabled, then an obvious criterion would be the change in the re-employment probability attributable to this programme. The evaluation methods described in this contribution are mainly established on such quantitative measures. Whereas evaluation in the US, for example, has mainly focused on earnings after training participation, in Europe the training effect on employment plays a dominant role. This is not surprising, as unemployment in Europe is much higher and more persistent than in the US. But the same quantitative outcome, say improvement of employment situation, may plausibly be measured in several ways, for instance by hours per week in the new job or by a simple distinction between employment and unemployment. Schmidt (1999) notes, that there are some problems arising with the choice of an appropriate outcome measure. Outcomes may not be comparable across interventions. Therefore a policy-maker who has to decide which measure to implement will normally try to translate the gains of a programme into monetary terms or to carry out a so-called cost-utility analysis. Still this is not easy because new problems like time or group preferences emerge. We will pick up these problems in the following section. Trying to measure the gains of a social programme using monetary terms, however, does not mean that other more qualitative aims and objectives, such as the quality of life of the affected participants or fairness aspects, are of less importance. Other

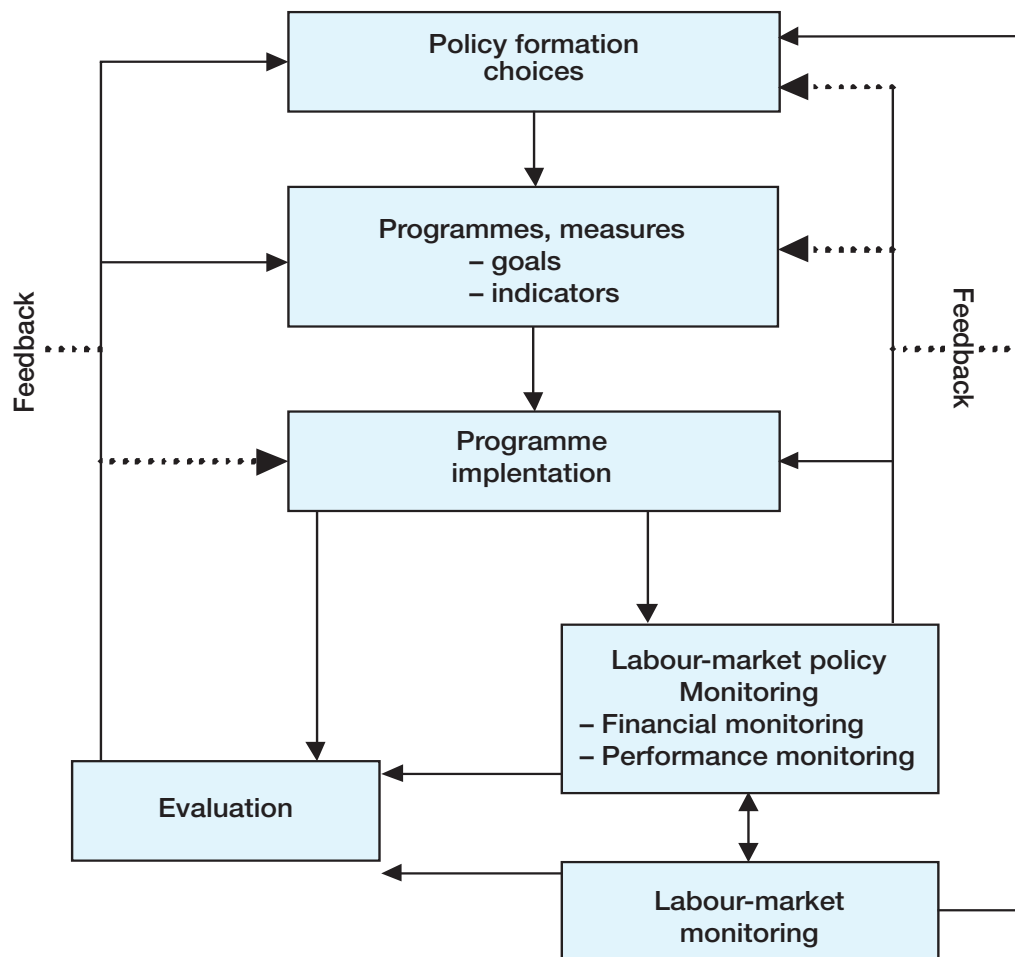
important examples for more qualitative aims and objectives include equity goals, social inclusion, civic participation or reduction in crime. We will address these issues later on.

2.1.4. The evaluation method

Having defined the programme and the success criteria, evaluating the impact of a programme requires disentangling the effect of the programme from other exogenous factors. The programme can be seen as the independent variable, whereas the dependent variable is the success criterion. Various empirical strategies can be used to measure the impact of the independent variable. The most favoured empirical strategy is a natural experiment where individuals participating in social programmes and those who were excluded from participation are randomly selected so that differ-

ences in the outcome variable after the treatment, i.e. after the programme took place, are solely attributable to the programme. Although seen as the golden path in evaluation, experiments have their own drawbacks, the most severe being an ethical one. It is simply very hard to refuse help to some people who are supposed to be in need of it. Besides this ethical issue, experiments suffer also from other shortcomings, such as randomisation bias, Hawthorne effect, disruption and substitution bias. We will address these issues later on. Quasi-experimental strategies, on the other hand, rely on a non-randomly chosen group of non-participants and differ in the way the control group is constructed. Examples include the before-after-estimator or the matching-procedure. This paper will address issues concerned with these estimators.

Figure 2: Major aspects of an evaluation process



Following Auer and Kruppe (1996) the major aspects which have to be taken into account when conducting an evaluation of social programmes are once again summarised in Figure 2. With the definition of the goals and objectives of the programme, appropriate indicators for measuring the success of the programme should be defined as well. These goals are then, once the programme has been implemented, confronted with the actual data obtained from a first monitoring. The following sections will mainly focus on the last aspect, i.e. derive some suggestions and advice on which econometric evaluation techniques should be used under different economic circumstances in order to conduct the evaluation.

2.2. Microeconomic evaluation

2.2.1. Fundamental evaluation problem

Inference about the impact of a treatment on the outcome of an individual involves speculation about how this individual would have performed in the labour market, had he or she not received the treatment ⁽⁵⁾. The framework serving as a guideline for the empirical analysis of this problem is the potential outcome approach, also known as the Roy-Rubin-model (Roy, 1951) (Rubin, 1974). In the basic model there are two potential outcomes (or responses), Y^1 and Y^0 , for each individual, where Y^1 indicates an outcome with training and without. In the former case, the individual is in the treatment group and in the latter case it is in the comparison group. To complete the notation we define a binary assignment indicator D , indicating whether an individual actually participated in training ($D=1$) or not ($D=0$). The treatment effect for each individual is then defined as the difference between his/her potential outcomes:

$$\Delta = Y^1 - Y^0. \quad (1)$$

The fundamental problem of evaluating this individual treatment effect arises because the observed outcome for each individual is given by:

$$Y = D \cdot Y^1 + (1 - D) \cdot Y^0. \quad (2)$$

This means that for individuals who participated in training ($D = 1$) we observe Y^1 and for those who did not participate we observe Y^0 . Unfortunately, we can never observe Y^1 and Y^0 for the same individual simultaneously and therefore we cannot estimate (1) directly. The unobservable component in (1) is called the counterfactual outcome, so that for an individual who participated in the training measure ($D = 1$), Y^0 is the counterfactual outcome, and for another one who did not participate it is Y^1 .

The concentration on a single individual requires that the effect of the intervention on each individual is not affected by the participation decision of any other individual, i.e. the treatment effect Δ for each person is independent of the treatment of other individuals. In statistical literature (Rubin, 1980) this is referred to as the stable unit treatment value assumption (SUTVA) and guarantees that average treatment effects can be estimated independently of the size and composition of the treatment population ⁽⁶⁾. Note that there will never be an opportunity to estimate individual effects with confidence. Therefore we have to concentrate on the population average of gains from treatment. The most prominent evaluation parameter is the so-called mean effect of treatment on the treated:

$$E(\Delta | D = 1) = E(Y^1 | D = 1) - E(Y^0 | D = 1). \quad (7)$$

The expected value of the treatment effect Δ is defined as the difference between the expected values of the outcome with and without training for those who actually participated in training. In the sense that this parameter focuses directly on actual training participants, it determines the realised gross gain from the training programme and can be compared with its costs, helping to decide whether the programme is a success or not (Heckman et al., 1997, 1998b; Heckman et al., 1999).

⁽⁵⁾ This is clearly different from asking whether there is an empirical association between training and the outcome (Lechner, 2000). See Holland (1986) for an extensive discussion of concepts of causality in statistics, econometrics and other fields.

⁽⁶⁾ Among other things SUTVA excludes cross-effects or general equilibrium effects. Its validity facilitates a manageable formal setup; nevertheless in practical applications it is frequently questionable whether it holds.

⁽⁷⁾ E is the expectation operator, in that case the expected value of the treatment effect Δ . $E(\Delta | D = 1)$ is the expected treatment for those who participated ($|$ means conditional on)

Despite the fact that most evaluation research focuses on average outcomes, partly because most statistical techniques focus on mean effects, there is also a growing interest regarding effects of policy variables on distributional outcomes. Examples where distributional consequences matter for welfare analysis include subsidised training programmes (LaLonde, 1995) or minimum wages (DiNardo et al., 1996). Koenker and Biliias (2002) show that quantile regression methods can play a constructive role in the analysis of duration (survival) data too. They describe the link between quantile regression and the transformation model formulation of survival analysis, offering a more flexible analysis than conventional methods, for example if one is interested in the duration of employment/unemployment after the programme took place.

Nevertheless we will focus on the average treatment effect on the treated $E(\Delta \mid D = 1)$ in this paper. The second term on the right side in equation (3) is unobservable as it describes the hypothetical outcome without treatment for participants in a programme. If the condition:

$$E(Y^0 \mid D = 1) = E(Y^0 \mid D = 0) \quad (4)$$

holds, we can use the non-participants as an adequate control group. In other words we would take the mean outcome of non-participants as a proxy for the counterfactual outcome of participants. This identifying assumption is definitely valid in social experiments. The key concept here is the randomised assignment of individuals into treatment and control groups. Individuals who are eligible to participate, for example, in training are randomly assigned to a treatment group that participates in the programme and a control group that does not. This assignment mechanism is a process that is completely beyond employee or administrator control. If the sample size is sufficiently large, randomisation will generate a complete balancing of all relevant observable and unobservable characteristics across treatment and control groups. Therefore the comparability between experimental treatment and control groups is facilitated enormously. On average, the two groups do not systematically differ except for

having participated in training. As a result, any observed difference in the outcomes of the groups after training is supposed to be solely induced by the programme itself, i.e. the impact of training is isolated and there should be no selection bias. Formally, random assignment ensures that the potential outcomes are independent of the assignment to the training programme. We write:

$$Y^1, Y^0 \perp\!\!\!\perp D \quad (5)$$

$\perp\!\!\!\perp$ denoting independence. When assignment to treatment is completely random it follows that:

$$E(Y^1 \mid D = 1) = E(Y^1 \mid D = 0),$$

and:

$$E(Y^0 \mid D = 1) = E(Y^0 \mid D = 0) \quad (6)$$

Therefore, treatment assignment becomes ignorable (Rubin, 1974) and we get an unbiased estimate of $E(\Delta)$, i.e. the randomly generated group of non-participants can be used as an adequate control group to estimate consistently the counterfactual term $E(Y^0 \mid D = 0)$ and thus the causal training effect $E(\Delta \mid D = 1)$. Although this approach seems to be very appealing in providing a simple solution to the fundamental evaluation problem, there are also problems associated with it. Besides relatively high costs and ethical issues concerning the use of experiments, in practice, a randomised experiment may suffer from similar problems that affect behavioural studies. Bijwaard and Ridder (2000) investigate the problem of non-compliance to the assigned intervention, that is when members of the treatment sample drop out of the programme and members of the control group participate. If the non-compliance is selective, i.e. correlated with the outcome variable, the difference of the average outcomes is a biased estimate of the effect of the intervention, and correction methods have to be applied too. Besides relatively high costs and ethical issues concerning the use of experiments, further methodological problems might arise, such as substitution or randomisation bias, which make the use of experiments questionable ⁽⁸⁾. For an extensive discussion of

⁽⁸⁾ A randomisation bias occurs when random assignment causes the types of persons participating in a programme to differ from the type that would participate in the programme as it normally operates, leading to an unrepresentative sample. We talk about a substitution bias when members of an experimental control group gain access to close substitutes for the experimental treatment (Heckman and Smith, 1995).

these topics the interested reader should refer to Burtless (1995), Burtless and Orr (1986) and Heckman and Smith (1995) ⁽⁹⁾.

More important for practical applications is the fact that, in most European countries, experiments are not conducted and researchers have to work with non-experimental data anyway. In non-experimental data, equation (4) will normally not hold:

$$E(Y^0 | D = 1) \neq E(Y^0 | D = 0), \quad (7)$$

The use of the non-participants as a control group might, therefore, lead to a selection bias. Heckman and Hotz (1989) point out that selection might occur on observable or unobservable characteristics. Good examples for observable characteristics are sociodemographic variables like qualification, age or gender of the individual, which are usually available in evaluation datasets. Unobservable characteristics might be the motivation or working habits of an individual. The aim of any observational evaluation approach is to ensure the comparability of treatment and control group by design; that is through a plausible identifying assumption. Taking account of observable factors might not be sufficient, if unobservable factors invalidate the comparison, for example when more motivated workers have a higher employment probability and are also more likely to participate in a training programme (Schmidt, 1999) ⁽¹⁰⁾.

In the following subsections we will present four different evaluation approaches. Each approach invokes different identifying assumptions to construct the required counterfactual outcome. Therefore each estimator is only consistent in a certain restrictive environment. As Heckman et al. (1999) note, all estimators would identify the same parameter only if there is no selection bias at all.

2.2.2. Before-after estimator

The most obvious and still widely used evaluation strategy is the before-after estimator (BAE). It compares the outcome of participants before training took place with their outcome after training. The basic idea is that the observable outcome in the pre-training period t' represents a valid description of the unobservable counterfactual outcome of the participants without training in the post-training period t . The central identifying assumption of the BAE can be stated as:

$$E(Y_t^0 | D = 1) = E(Y_{t'}^0 | D = 1). \quad (8)$$

Given the identifying assumption in (8), the following estimator of the mean treatment effect on the treated can be derived:

$$\Delta^{BAE} = E[(Y_t^1 | D = 1) - (Y_{t'}^0 | D = 1)]. \quad (9)$$

Heckman et al. (1999) note that conditioning on observable characteristic X makes it more likely that assumption (8) will hold. If the distribution of X characteristics is different between the treatment and the control group, conditioning on X may eliminate systematic differences in the outcomes ⁽¹¹⁾.

The validity of (9) depends on a set of implicit assumptions. First, the pre-exposure potential outcome without training should not be affected by training. This may be invalid if individuals have to behave in a certain way in order to get into the programme or behave differently in anticipation of future training participation. Second, no time-variant effects should influence the potential outcomes from one period to the other. If there are changes in the overall state of the economy or changes in the lifecycle position of a cohort of participants, assumption (8) may be violated (Heckman et al., 1999).

A good example where this might be the case is Ashenfelter's dip. That is a situation where shortly

⁽⁹⁾ As Smith (2000) notes, social experiments have become the method of choice in the US. The most famous among them is the National Job Training Partnership Act which had a major influence regarding the view on non-experimental studies. In Europe, however, social experiments have not received similar acceptance, although recently some test experiments have been conducted. The most important one is the Restart experiment in Britain (Dolton and O'Neill, 1996).

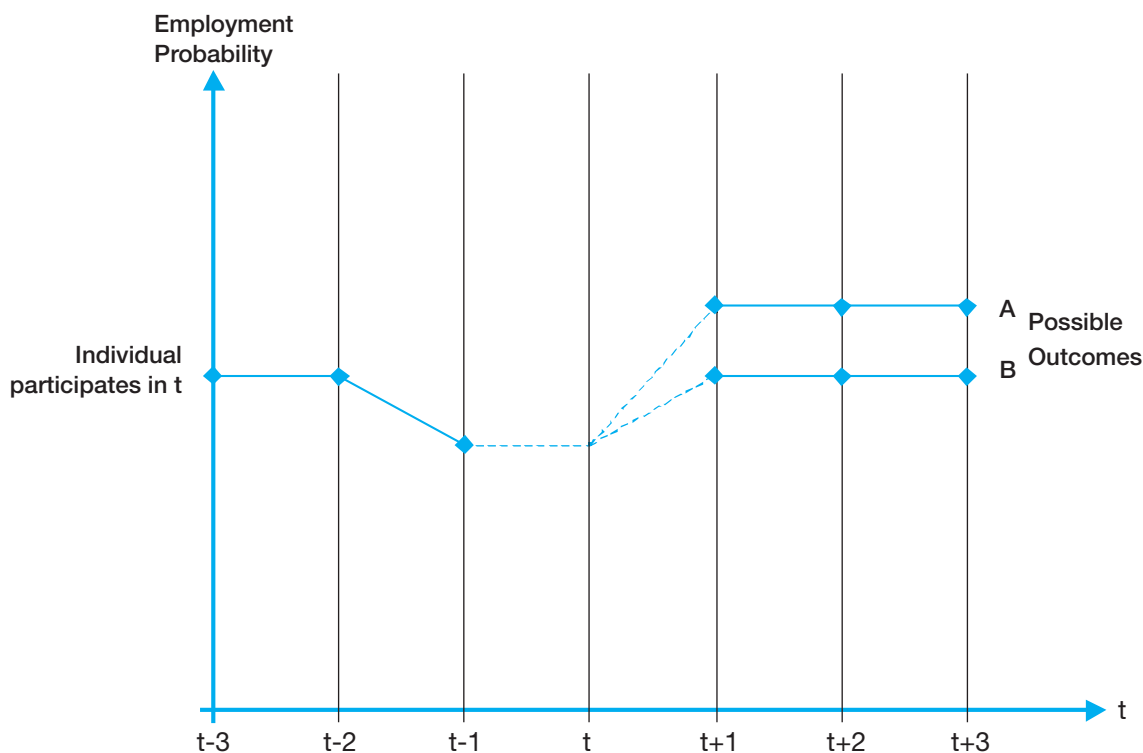
⁽¹⁰⁾ The distinction between observable and observed factors depends on the dataset. Whereas some factors like the motivation of an individual are hard to measure and therefore usually not observable, other factors like education are in general observable but might not be observed in the dataset at hand.

⁽¹¹⁾ On the other hand, if the difference in the treatment and the control group are due to unobservable characteristics, conditioning may accentuate rather than eliminate the differences in the no-programme state between the both groups (Heckman et al., 1999).

before the participation in an ALMP programme, the employment situation of the future participants deteriorates. Ashenfelter (1978) found this dip while evaluating the effects of treatment on earnings, but later research demonstrated that this dip can be observed on employment probabilities for

participants too. If the dip is transitory and only experienced by participants, assumption (8) will not hold. By contrast, permanent dips are not problematic, as they affect employment probability before and after treatment in the same way. Figure 3 illustrates Ashenfelter's dip.

Figure 3: Ashenfelter's dip



We assume that an individual participates in a training programme in period t and experiences a transitory dip in the pre-training period $t - 1$; for example the employment probability may be lowered as the individual is not actively seeking work because of the intended participation in a programme in the next period. There are no economy-wide, time-varying effects that influence the employment probability of the individual. After the programme takes place, employment probability might assume many values. For the sake of simplicity we consider two cases. Case A assumes that there has been a positive effect on employment probability (vertical difference between A and B). As the BAE compares the employment probability of the individual in period

$t + 1$ and $t - 1$ (vertical difference between point A and C) it would overestimate the true treatment effect, because it would attribute the restoration of the transitory dip completely to the programme. In a second example, we assume that there has been no treatment effect at all (B). In period $t + 1$ the employment probability is restored to its original value before the dip took place in $t - 1$. Again the BAE attributes this restoration completely to the programme and would estimate a positive treatment effect (vertical difference between point B and C). Clearly the problem of Ashenfelter's dip could be avoided, if a time period is chosen as a reference level before the dip took place, for example period $t - 2$. But as we will not know *a-priori* in an

empirical application when the dip starts, the question arises of which period to choose.

A major advantage of the BAE is that it does not require information on non-participants. All that is needed is longitudinal data on outcomes on participants before and after the programme took place ⁽¹²⁾. As the employment status of the participants is usually known in the programmes or even prerequisite for participation, the BAE does not impose any major problems regarding the data availability, which might explain why it is still widely used ⁽¹³⁾.

2.2.3. Cross-section estimator

The basic idea of the cross-section estimator is to compare the outcome of participants after the programme took place with the outcome of non-participants at the same time period. Instead of comparing participants at two different time periods, the cross-section estimator compares participants and non-participants at the same time (after the programme took place), i.e. the population average of the observed outcome of non-participants replaces the population average of the unobservable outcome of participants. This is useful if no longitudinal information on participants is available or macroeconomic conditions shift substantially over time (Schmidt, 1999). The identifying assumption of the cross-section estimator can be stated formally as:

$$E(Y_t^0 | D=1) = E(Y_t^0 | D=0). \quad (10)$$

That is, those who participate in the programme have, on average, the same no-treatment outcome as those who do not participate. If this assumption is valid, the following estimator of the mean true treatment effect can be derived:

$$\Delta^{CSE} = E[(Y_t^1 | D=1) - (Y_t^0 | D=0)]. \quad (11)$$

It is worth noting that conditioning on observable characteristics makes it more likely that this assumption will hold. If the distribution of X characteristics is different between the treatment and the control group, conditioning on X may elimi-

nate systematic differences in the outcomes ⁽¹⁴⁾. To give a straightforward example, let us assume that X represents the qualification level of an individual. For the sake of simplicity, we assume that X might only take two values (1 for high-skilled and 0 for low-skilled workers). Therefore, conditioning on X results in estimating the treatment effects separately for both skill groups and is intuitively appealing. The identifying assumption is then:

$$E(Y_t^0 | X, D=1) = E(Y_t^0 | X, D=0). \quad (12)$$

The first approach in equation (11) can be seen as a 'naive' estimator, because it just compares the results for the whole group, whereas the second takes into account observed differences in individual's characteristics, such as different skill levels. The resulting estimator can be written as:

$$\Delta^{CSE|X} = E[(Y_t^1 | X, D=1) - (Y_t^0 | X, D=0)]. \quad (13)$$

Schmidt (1999) notes, that for assumption (12) to be valid, selection into treatment has to be statistically independent of its effects given X (exogenous selection), that is, no unobservable factor should lead individual workers to participate. A good example where this is violated might be the case if motivation plays a role in determining the desire to participate and the outcomes without treatment. Then we have, even in the absence of any treatment effect, a higher average outcome in the participating group compared to the non-participating group. Ashenfelter's dip is not problematic for the cross section estimator as we compare only participants and non-participants after the programme took place. Moreover, as long as economy-wide shocks and individual lifecycle patterns operate identically for the treatment and the control group, the cross-section estimator is not vulnerable to the problems that plague the BAE (Heckman et al., 1999).

2.2.4. Matching estimator

The matching approach originated in statistical literature and shows a close link to the experi-

⁽¹²⁾ The BAE might also work with repeated cross-sectional data from the same population, not necessarily containing information on the same individuals. See Heckman and Robb (1985) or Heckman et al. (1999) for details.

⁽¹³⁾ Note that the labour-market status (unemployed, part-time or low-skilled employed, etc.) is very often one of the entry conditions for ALMP programmes.

⁽¹⁴⁾ On the other hand, if the differences in the treatment and the control group are due to unobservable characteristics, conditioning may accentuate rather than eliminate the differences in the no-programme state between both groups (Heckman et al., 1999).

mental context ⁽¹⁵⁾. The basic idea underlying the matching approach is to find in a large group of non-participants those individuals who are similar to the participants in all relevant pre-training characteristics. That being done, the differences in the outcomes between the well selected and thus adequate control group and the trainees can then be attributed to the programme. Matching does not need to rely on functional form or distributional assumptions, as its nature is non-parametric (Augurzky, 2000).

Matching is first of all plagued by the same problem as all non-experimental estimators, which means that assumption (4) cannot be expected to hold when treatment assignment is not random. However, following Rubin (1977), treatment assignment may be randomly given a set of covariates. The construction of a valid control group via matching is based on the identifying assumption that conditional on all relevant pre-training covariates Z , the potential outcomes (Y^1, Y^0) are independent of the assignment to training ⁽¹⁶⁾. This so-called conditional independence assumption (CIA) can be written formally as:

$$Y^1, Y^0 \perp\!\!\!\perp D \mid Z. \quad (17)$$

If assumption (14) is fulfilled we get:

$$E(Y^0 \mid Z, D = 1) = E(Y^0 \mid Z, D = 0) = E(Y^0 \mid Z). \quad (15)$$

Similar to randomisation in a classical experiment, the role of matching is to balance the distributions of all relevant pre-treatment characteristics in the treatment and control group, and thus to achieve independence between potential outcomes and the assignment to treatment, resulting in an unbiased estimate. The exact matching estimator can be written as:

$$\Delta^{MAT} = E[(Y_i^1 \mid Z, D = 1) - (Y_i^0 \mid Z, D = 0)]. \quad (16)$$

Conditioning on all relevant covariates is, however, limited in case of a high dimensional vector Z . For instance, if Z contains n covariates

which are all dichotomous, the number of possible matches will be 2^n . In this case cell matching, that is exact matching on Z , is not possible since an increase in the number of variables increases the number of matching cells exponentially. To deal with this dimensionality problem, Rosenbaum and Rubin (1983) suggest the use of balancing scores $b(Z)$, i.e. functions of the relevant observed covariates Z such that the conditional distribution of Z given $b(Z)$ is independent of the assignment to treatment, that is $Z \perp\!\!\!\perp D \mid b(Z)$ holds.

For trainees and non-trainees with the same balancing score, the distributions of the covariates Z are the same, i.e. they are balanced across the groups. Moreover Rosenbaum and Rubin (1983) show that if the treatment assignment is strongly ignorable ⁽¹⁸⁾ when Z is given, it is also strongly ignorable given any balancing score. The propensity score, i.e. the probability of participating in a programme is one possible balancing score. It summarises the information of the observed covariates into a single index function.

Rosenbaum and Rubin (1983) show how the conditional independence assumption extends to the use of the propensity score so that:

$$Y^0 \perp\!\!\!\perp D \mid P(Z) \quad (17)$$

Therefore we get:

$$E(Y^0 \mid P(Z), D = 1) = E(Y^0 \mid P(Z), D = 0) = E(Y^0 \mid P(Z)), \quad (18)$$

which allows us to rewrite the crucial term in the average treatment effect (3) as:

$$E(Y^0 \mid D = 1) = E_{p(z)}[(Y^0 \mid P(Z), D = 0) \mid D = 1]. \quad (19)$$

Hujer and Wellner (2000b) note that the outer expectation is taken over the distribution of the propensity score in the treated population. The major advantage of the identifying assumption (17) is that it transforms the estimation problem into a much easier task since one has to condition on a univariate scale, i.e. on the propensity

⁽¹⁵⁾ See Rubin (1974, 1977, 1979), Rosenbaum and Rubin (1983, 1985a, 1985b) or Lechner (1998).

⁽¹⁶⁾ If we say relevant we mean all those covariates that influence the assignment to treatment as well as the potential outcomes. In contrast to the cross-section estimator, the matching procedure can also use information from the pre-treatment period, such as employment status or other time-varying covariates. To make this difference clear, we denote the covariates by Z .

⁽¹⁷⁾ For the purpose of estimating the mean effect of treatment on the treated, the assumption of conditional independence of Y^0 is sufficient because we like to infer estimates of Y^0 for persons with $D = 1$ from data on persons with $D = 0$ (Heckman et al., 1997).

⁽¹⁸⁾ Strongly ignorable means that assumption (16) holds and: $0 < P(Z) = P(D = 1 \mid Z) < 1$. The latter ensures that there are no characteristics in Z for which the propensity score is zero or one. Proofs go beyond the scope of this work and can be found in Rosenbaum and Rubin (1983).

score, only. When $P(Z)$ is known, the problem of dimensionality can be eliminated. The evaluation of the counterfactual term via matching on the basis of the group of non-participants then only requires to pair participants with non-participants which have the same propensity score. This ensures a balanced distribution of Z across both groups. Unfortunately, $P(Z)$ will not be known *a-priori* so it has to be replaced by an estimate. This can be achieved by any number of standard probability models, for example a probit model.

The empirical power of matching to reduce the problem of selection bias relies crucially on the quality of the estimate of the propensity score, on the one hand, and on the existence of comparison persons that have equal propensity scores as the treated persons. If the latter is not ensured we run the risk of incomplete matching with biased estimates ⁽¹⁹⁾. Several procedures for matching on the propensity score have been suggested and will be discussed briefly; a good overview can be found in Heckman et al. (1998a) and Smith and Todd (2000). To introduce them a more general notation is needed. We estimate the effect of treatment for each observation in the treatment group, by contrasting his/her outcome with treatment with a weighted average of control group observations j in the following way:

$$Y_i^1 - \sum_{j \in \{D=0\}} W_{N_0 N_1}(i, j) Y_j^0, \quad (20)$$

where N_0 is the number of observations in the control group and N_1 is the number of observations in the treatment group and is a weighting function which can take different forms. Matching estimators differ in the weights attached to the members of the comparison group (Heckman et al., 1998a). Nearest neighbour (NN) matching sets:

$$C(P_i) = \min_j \|P_i - P_j\|, j \in N_0. \quad (21)$$

Doing so, the non-participant with the value of that is closest to is selected as the match, therefore $W_{N_0 N_1}(i, j) = 1$ for this unit and $W_{N_0 N_1}(i, j) = 0$ otherwise ⁽²⁰⁾. Several variants of nearest neighbour matching are proposed, for example nearest

neighbour matching 'with' and 'without replacement'. In the former case, a non-participating individual can be used more than once as a match, whereas in the latter case it is considered only once. Use of more than one nearest neighbour (over-sampling) is also suggested. Nearest neighbour matching runs the risk of bad matches, if the closest neighbour is far away.

This can be avoided by imposing a tolerance on the maximum distance $\|P_i - P_j\|$ allowed. This form of matching, caliper matching (Cochrane and Rubin, 1973), imposes the condition:

$$\|P_i - P_j\| < \varepsilon, j \in N_0. \quad (22)$$

where ε is a pre-specified level of tolerance.

Kernel matching (KM) is a non-parametric matching estimator that uses all individuals in the control group to construct a match for each programme participant. KM defines the weighting function as:

$$W_{N_0}(i, j) = \frac{K_{ij}}{\sum_{k \in \{D=0\}} K_{ik}}, \quad (23)$$

where $K_{ik} = K((P_i - P_k)/h)$ is a kernel that down-weights distant observations from P_i and h is a bandwidth parameter (Heckman et al., 1998a).

The matching estimator is very data-demanding in the sense that we need information for participants and non-participants before and after the programme took place. In the case of exact matching, a 'rich' dataset is needed to ensure that we find comparable individuals in the control group for every combination of observable characteristics. Even if we do not use exact matching but matching over the propensity score, a rich dataset is needed. In that case the quality of the score depends on our ability to account for all relevant covariates that determine the participation decision. Problems arise, if a programme is compulsory, for example if all the unemployed have to participate in a training programme at a certain point in their unemployment spell. In this case it might get difficult to find a suitable control group,

⁽¹⁹⁾ Matching was much discussed in recent econometric literature. Heckman and his colleagues reconsidered and further developed the identifying assumptions of matching stated by Rubin (1977) and Rosenbaum and Rubin (1983). It turns out that the new identifying assumptions are weaker compared to the original statements which brings along some advantages. Presenting these ideas goes beyond the scope of this work. The interested reader should refer to Heckman et al. (1996, 1997, 1998a and b) and Heckman and Smith (1995).

⁽²⁰⁾ Exact matching imposes an even stronger condition, where only non-participants with exactly the same propensity score or the same realisation of characteristics X are considered as matches.

because all the unemployed will be in the programme at some time. However, it is still possible to evaluate the optimal timing of a programme, for example after three, six or 12 months of unemployment.

2.2.5. Difference-in-differences estimator

It has been claimed that controlling for selection on observable characteristics may not be sufficient since remaining unobservable differences may still lead to a biased estimation of treatment effects. These differences may arise from differences in the benefits which individuals expect from participation in a treatment, which might influence their decision to participate. Furthermore, some groups might exhibit bad labour-market prospects or differences in motivation. These features are unobservable to a researcher and might cause a selection bias.

To account for selection on unobservables, Heckman et al. (1999) suggest econometric selection models and difference-in-differences (DID) estimators. The DID estimator requires access to longitudinal data and can be seen as an extension to the classical BAE. Whereas the BAE compares the outcomes of participants after they participate in the programme with their outcomes before they participate, the DID estimator eliminates common time trends by subtracting the before-after change in non-participant outcomes from the before-after change for participant outcomes. The simplest application of the method does not condition on X and forms simple averages over the group of participants and non-participants. Changes in the outcome variable Y for the treated individuals are contrasted with the corresponding changes for non-treated individuals (Heckman et al., 1998a):

$$\Delta^{DID} = [Y_t^1 - Y_t^0 | D=1] - [Y_t^0 - Y_t^0 | D=0]. \quad (24)$$

The DID estimator is based on the assumption of time-invariant linear selection effects. The critical identifying assumption of this method is that the biases are the same, on average, in different time periods before and after the period of participation in the programme, so that differencing the differences between participants and non-participants eliminates the bias (Heckman et al., 1998a).

To make this point clear, we assume that we want to evaluate a training programme that aims to improve the employment prospects of individuals. And let us further assume that the participants have a higher motivation and, owing to this higher motivation, the outcome for participants is on average 5% higher than for non-participants. If we additionally presume that this bias remains constant over time, we can estimate the effect of a programme by comparing the pre- and post-programme differences of the participants and non-participants. Let us say that the participants had an average outcome of 50% in the period before the programme took place, that the treatment effect is 10% and, owing to a cyclical upswing, the outcome is increased by 15%. Therefore the participants have an average outcome of 75% in the period after the programme took place. The non-participants have an average outcome of 45% in the first period (5% less than the participants because of the motivational selection effect) and 60% in the second period. Comparing the average outcomes for participants in period two (75%-60%=15%) is misleading since the selection effect due to unobservable characteristics is not taken into account. The DID estimator, in contrast, leads to the correct result of 10% ⁽²¹⁾. Through the differencing procedure, the cyclical upswing is correctly not attributed to the programme and neither is the motivational effect.

This example can be formally explained by denoting the outcome for an individual at time t as:

$$Y_{it} = \alpha_{it} + D_{it} \cdot Y_{it}^1 + (1 - D_{it}) \cdot Y_{it}^0, \quad (25)$$

where α_{it} captures the effects of selection on unobservables. The validity of the DID estimator relies crucially on the assumption:

$$\alpha_{it} = \alpha_{it}^0. \quad (26)$$

Only if the selection effect is time-invariant will it be cancelled out and an unbiased estimate result. The differencing leads to:

$$\begin{aligned} Y_{it} - Y_{it}^0 &= [D_{it} \cdot Y_{it}^1 + (1 - D_{it}) \cdot Y_{it}^0] - \\ &\quad [D_{it}^0 \cdot Y_{it}^1 + (1 - D_{it}^0) \cdot Y_{it}^0] + \\ &\quad [\alpha_{it} - \alpha_{it}^0]. \end{aligned} \quad (27)$$

⁽²¹⁾ Difference for participants: (75%-50%) = 25%; difference for non-participants: (60%-45%) = 15%; DID estimator: 25%-15% = 10%.

If (26) is fulfilled, the last term in the expression can be cancelled out, leading to an unbiased estimate. Compared to the method of matching, the DID approach does not require that the bias vanishes for any ‘matched’ individuals, but only that it remains constant (Heckman et al., 1998a).

If we condition the DID approach on observable characteristics X , the new estimator is given by:

$$\Delta^{DID|X} = [Y_t^1 - Y_t^0 | X, D=1] - [Y_t^0 - Y_t^0 | X, D=0]. \quad (28)$$

The identifying assumption of this method is:

$$E(Y_t^1 - Y_t^0 | X, D=1) = E(Y_t^0 - Y_t^0 | X, D=0). \quad (29)$$

Ashenfelter’s dip is definitely a problem for the DID estimator. If the dip is transitory and is eventually restored even in the absence of participation in the programme, the bias will not average out. Therefore Bergmann et al. (2000) apply a combination of the matching- and the DID estimator in a recent paper, by implementing a ‘conditional DID estimator’, where conditional means that treatment and control groups are already partly comparable regarding their observable characteristics. Kluve et al. (1999) suggest using the pre-treatment (labour market) histories of the individuals as an important variable in the matching process, so that only individuals with identical pre-treatment histories are compared. Basically the conditional DID estimator extends the one suggested from Heckman et al. (1998a) by adding a longitudinal dimension. The basic question in this case is deciding the appropriate time span to be considered.

2.2.6. Duration models

An alternative approach to modelling treatment effects of ALMP is the application of duration models. The basic idea here is to include the duration of (un)employment and the duration of the programmes when estimating effects. In a bivariate duration model the durations T_u and T_p measure the duration of unemployment until entry into employment and the duration until entry into ALMP, respectively. T_u and T_p are random variables, t_u and t_p denote their realisations. We are interested in the causal effect of participation in ALMP on exit from unemployment, that is the effect of the realisation of T_p on the distribution of T_u . This implies that the causal effect is captured

by the effect of t_p on the hazard rate for job finding: $\theta_u(t|T_p, x, v)$ for $t > t_p$.

The transition rate from unemployment to employment at time conditional on and can be specified by a mixed proportional hazard model as follows:

$$\theta_u(t|x, t_p, d_z, v_u) = \lambda_u(t) \cdot \exp(x' \beta_u + \delta(t|t_p, x) \cdot I(t > t_p) + \mu_{uz} \cdot d_z + v_u), \quad (30)$$

where $I(\cdot)$ denotes the indicator function, which is 1 if its argument is true and 0 otherwise. The function $\lambda_u(t)$ is called the ‘baseline hazard’ individual duration dependence. δ measures the effect of the participation in ALMP on the transition rate from unemployment to employment, x is a vector of explanatory variables, μ_{uz} measures whether there is any benefit exhaustion effect and the term d_u represents unobserved heterogeneity. In a similar way, the hazard rate to programme p at time t conditional on x and d_z can be specified in the following equation:

$$\theta_p(t|x, d_z) = \lambda_p(t) \cdot \exp(x' \beta_p + \mu_{pz} d_z + v_p). \quad (31)$$

In the equation for θ_u and θ_p , unobserved heterogeneity is allowed to affect the transitions to both a job and to a programme.

‘If the unobserved characteristics have a negative effect on the job-finding rate and a positive effect on the transition to a programme, then conditional on the observed characteristics and the elapsed duration of unemployment, the average quality of workers in a programme is lower than the average quality of workers who do not enter a programme. Then, if we would simply compare the transition rates to regular jobs of both groups we would compare workers with unfavourable characteristics and programme participation with workers with more favourable characteristics and non-participation. Therefore, we would underestimate the true effect of participating in a programme. The opposite effect is also possible. One could imagine that the people in control of the programmes want their programmes to be a success. Therefore they prefer workers with good characteristics to flow into their programme. This would imply that there is a positive correlation between unobserved heterogeneity components in both transition rates. Then we would overestimate the treatment effect of programmes.’ (Lalive et al., 2000).

The authors of the empirical studies assume for the joint distribution of the unobserved characteristics $G(v_u, v_p)$ a multivariate discrete distribution using a multinomial logit specification (Heckman and Singer, 1984).

2.2.7. An intuitive example

Table 1 summarises the different evaluation estimators introduced in the previous subsections. It presents short descriptions of the various approaches and discusses their advantages and disadvantages using a simple numerical example ⁽²²⁾. The goal of this example is not to show the (dis-) advantages of each evaluation estimator in detail, but to give some guidance on

how they perform in a certain economic context. For the sake of simplicity it is assumed that only two types of workers (highly-skilled ($X=1$) and low-skilled ($X=0$)) exist. It is further assumed that the interest lies in evaluating a training programme that is intended to improve the skills of the workers and therefore enhance their employment prospects. To make the importance of heterogeneity clear it is furthermore assumed that the programme works better for highly-skilled workers. It can be thought of as a very specialised training measure to improve, for example, management skills and is therefore only taken up by a relatively small group of highly-skilled workers.

Table 1: **Comparison of the different evaluation estimators**

Estimator	Description	(Dis-) Advantages
BAE	Compares the outcome for participants before and after the programme took place.	+ easy to implement + low data requirements (no information for non-participants needed) - economic changes from the before to the after period might be falsely attributed to the programme - Ashenfelter's dip
CSE	Compares the outcome for participants in the period after the programme took place, with the outcome for non-participants in the same period.	+ economy-wide changes are not attributed to the programme + Ashenfelter's dip is not a problem - needs data for participants and non-participants for the period after the programme took place
Matching estimator	Compares the outcome of participants in the period after the programme took place, with the outcome of matched (statistical twins) non-participants in the same period.	+ takes account of selection on observable characteristics + Ashenfelter's dip is not a problem - needs data for participants and non-participants before and after the programme took place (e.g. labour-market history) - unobservable characteristics
DID estimator	Compares the before-after change for participant outcomes with the before-after change for non-participant outcomes.	+ takes account of selection on unobservable characteristics - needs data for participants and non-participants before and after the programme took place - Ashenfelter's dip
Duration models	The effects are measured by the coefficient of an explanatory dummy variable (participation yes/no) in a bivariate model framework.	+ takes into account spells (that is duration of (un)employment and policy measures) - selection problem is solved only implicitly

⁽²²⁾ See Schmidt (1999) for a similar example.

Table 2 displays the labour-market outcomes of the workers who received ($D=1$) and did not receive ($D=0$) treatment. Columns 2 and 3 show the labour-market outcomes for non-participants before (Y_{t-}) and after (Y_t) the programme took place. Column 5 contains the pre-programme outcome of participants (Y_{t-}), whereas column 7 contains the post-programme outcome of participants ($Y_{t+\Delta}$). All these outcomes are actually observed. Column 6 contains the counterfactual outcome under no treatment for participants. Clearly, this outcome is never observed in an

empirical study but allows us to answer the question ‘What would have happened to the participants if they had not participated?’. The programme impact can be calculated by comparing the actual outcome of the participants after the programme took place ($Y_{t+\Delta}$) with their counterfactual outcome under no treatment (Y_t). Again it is important to note that a researcher will not be able to do so since column 6 is unobservable. This unobservable outcome is only introduced as a reference level to illustrate the functioning of the different estimators.

Table 2: **Labour-market outcomes for participants and non-participants**

Non-Participants (D = 0)			Participants (D = 1)			
Low-skilled workers (X = 0)						
Worker	Y _{t'}	Y _t	Worker	Y _{t'}	Y _t	Y _t +Delta
1	0	1	11	0	0	1
2	0	0	12	0	0	0
3	0	0	13	0	0	0
4	0	0	14	0	1	1
5	0	0	15	0	1	1
6	1	1	16	0	0	0
7	1	1	17	0	1	1
8	1	1	18	1	1	1
9	1	1	19	1	1	1
10	1	1	20	1	0	1
Mean	0.5	0.6		0.3	0.5	0.7
Highly-skilled workers (X = 1)						
Worker	Y _{t'}	Y _t	Worker	Y _{t'}	Y _t	Y _t +Delta
21	0	0	31	0	0	1
22	0	1	32	0	1	1
23	0	0	33	0	0	1
24	1	1	34	1	1	1
25	1	1	35	1	1	1
26	1	1				
27	1	1				
28	1	1				
29	1	1				
30	1	1				
Mean	0.7	0.8		0.4	0.6	1

The two skill groups will be discussed separately.

In the group of low-skilled workers there are 10 non-participants (workers 1-10) and 10 participants (workers 11-20). 50 % of the non-participants are employed in both periods and for one non-participant the labour-market situation improves over time (worker 1). In the group of participants only 30 % are employed in the first period. For three workers (workers 14, 15 and 17) the situation would have improved even in absence of the training programme (Y_i), whereas for one worker (worker 20) it would have worsened. The programme impact can be estimated by comparing columns 7 and 6. The programme improves the employment prospects for two workers (workers 11 and 20) and therefore the effect is 0.2. The following is how a researcher would calculate the programme impacts with some of the presented estimators ⁽²³⁾.

The BAE invokes the identifying assumption that the pre-training outcome of the participants represents a valid description of their unobservable counterfactual outcome in the post-training period (equation 7). Therefore the estimator is calculated by comparing the outcome in column 7 (0.7) with the outcome in column 5 (0.3) and leads to a result of 0.4. Clearly, this estimator attributes all improvement in the employment situation between the two periods to the programme. But since, in our example, the situation for two workers would have improved anyway, for example due to an overall economic improvement, the BAE overestimates the real impact.

The implementation of the cross-section estimator requires that the population average of the observed outcome of the non-participants is equal to the population average of the unobservable outcome of participants (equation 10). Therefore the estimator is calculated by comparing the actual outcome for participants (workers 11-20) in period t (average: 0.7, column 7) with the actual outcome for non-participants (workers 1-10) in the same period (average: 0.6, column 3). The estimated impact

would be 0.1. This is due to the fact that the constructed labour-market situation for non-participants in t (average: 0.6, column 3) is very similar to the counterfactual labour-market situation for participants (average: 0.5, column 6).

A matching estimator replaces the counterfactual outcome of the participants with the population average of a matched control group. In this example there is only one available matching variable, namely the labour-market situation before training. Estimation is then simply done by comparing the outcome in t of workers 11-17 (average: 0.57, column 7) with the outcome of workers 1-5 (average: 0.2, column 3) on the one hand and between workers 18-20 (average: 1.0, column 7) and 6-10 (average: 1.0, column 3) on the other. The estimated impact is then a weighted average (by the number of participants) between both groups. The effect in the first group (seven participants unemployed before training) is 0.37, the effect in the second group is 0 (three workers employed before training). The weighted average is, therefore, $(0.37 \times 7/10) = 0.26$.

Finally, a DID estimator eliminates common time trends by subtracting the before-after change in non-participant outcomes from the before-after change for participant outcomes. In our example we form simple averages over the group of participants and non-participants and contrast changes in the labour-market situation for the treated individuals with the corresponding changes for non-treated individuals. The estimated impact in the example is then $(0.7 - 0.3) - (0.6 - 0.5) = 0.3$. Since the increase in potential non-treatment outcome is higher for participants (0.2 vs. 0.1), the DID overestimates the true treatment effect.

The same calculations can be done for the group of highly-skilled workers. This group contains 10 non-participants (workers 21-30) and five participants (workers 31-35). 60 % of the participants are unemployed before training and only one (worker 32) would have experienced an improvement in his labour-market situation without training. The programme effect in the group of highly-skilled workers is 0.4. The

⁽²³⁾ For the sake of simplicity this numerical example has been kept on a basic methodological level. Therefore, more complex estimators like the conditional DID or the duration models could not be illustrated and the focus lies on the BAE, CSE, DID and the exact matching estimator.

evaluation estimators can be calculated analogously as described above and the results can be found in column 3 of Table 3. The matching estimator is very successful this time as it

exactly calculates the true impact, whereas the BAE and DID overestimate the programme impact and the cross-section estimator underestimates it.

Table 3: **Estimates of the programme effects**

	Low-Skilled	Highly-Skilled	Average
Impact	0.20	0.40	0.27
BAE	0.40	0.60	0.47
CSE	0.10	0.20	0.13
Matching	0.26	0.40	0.31
DID	0.30	0.50	0.37

The last column is a weighted average of the programme effects in the two groups (weighted according to the number of participants). This is simply to show how important it is to account for heterogeneity in the estimation of programme impacts. Due to data limitations previous evaluation studies had to pool heterogeneous programmes and/or heterogeneous individuals and estimate one composite treatment effect. That this is a misleading approach becomes clear in the given example. The weighted impact for low- and highly-skilled workers is 0.27. Since the true impact for highly-skilled (low-skilled) workers is 0.4 (0.2), disregarding the heterogeneity of the worker leads to an underestimation (overestimation) of the true effect. Clearly this is an unnecessary source of bias which has to be avoided. Therefore, heterogeneity with respect to individual characteristics, regional aspects and, in particular, different programmes should be modelled in the evaluation.

2.3. Macroeconometric evaluation

In this chapter we will discuss the impact of ALMP, not on particular individuals but on aggregate economic variables. Instead of looking at the effect on individual performance we would like to know if the ALMP represent a net gain for the whole economy. If the total number of jobs is not affected by labour-market policies, the effects will be distributional only. This might be desirable, for example, if work is shifted from the old to the

young, but can hardly justify the substantial fiscal costs of the ALMP.

The need for macroeconometric evaluation results from the need to estimate whether a positive effect on the microeconomic level is also positive on the macroeconomic level. This is a question of spillover effects, i.e. if the effects on the participants is counteracted or enforced by the effects on the non-participants. ALMP is often thought to have a positive effect on the participants but negative effects on the non-participants. Important effects in this context are the so-called dead-weight losses and substitution effects that have received substantial attention in the literature (Layard et al., 1991 or OECD, 1993), mainly in the context of job creation schemes. If the programme outcome is no different from that in its absence, we talk about a dead-weight loss. A common example is the hiring from the target group that would have occurred with or without the programme. If a firm hires a subsidised worker instead of an unsubsidised worker we talk about a substitution effect. The net short-term employment effect in this case is zero. Calmfors (1994) defines 'the substitution effect as the extent to which jobs created for a certain category of workers simply replace jobs for other categories, because relative wage costs are changed.' Such effects are likely in the case of subsidies for private-sector work. There is always a risk that employers hold back ordinary job creation in order to be able to take advantage of the subsidies. In order to minimise this danger, a principle of additionality may be imposed.

Another problem might be that active labour-market programmes may crowd out

regular employment. This can be seen as a generalisation of the so-called displacement effect. This effect typically refers to displacement in the product market; for example, firms with subsidised workers may increase output, while output may be reduced in firms that do not have subsidised workers. Clearly these effects have to be taken into account before making statements about the net effect of ALMP.

To derive the empirical model for a macroeconomic evaluation, the theoretical analysis of ALMP becomes crucial. This is because the specification of the empirical model matters in addition to the outcome (i.e. dependent) variable. Calmfors (1994) shows how a theoretical framework can be developed to allow an analysis of ALMP. He presents the 'revised Layard-Nickell model' as a basic framework for the analysis of the effects of ALMP on a number of economic variables and processes that influence aggregate employment and unemployment rates. Therefore, such a framework can be used to identify the different effects of ALMP on the whole economy. Following the discussion of Calmfors (1994), important effects of ALMP are considered below.

The first effects relate to the matching process. ALMP can improve matching between workers and jobs through several channels. First, ALMP can improve the active search behaviour of participants. Second, ALMP can speed up the matching process by adjusting the structure of labour supply to demand. Here we primarily think of retraining programmes that adapt the skills of the unemployed to the requirements of vacant jobs. Third, participation in an ALMP programme can serve as a substitute for work experience that reduces the employer's uncertainty about the employability of the job applicant.

If ALMP can improve the matching process, what are the effects on regular employment or wages? First, an improved matching process means that for a given stock of vacancies there is a greater inflow into employment ⁽²⁴⁾. Furthermore, the improved matching process reduces the average duration that a vacancy remains unfilled. Since this reduces the costs of maintaining a vacancy, firms provide more vacancies, which is equivalent to an increase in labour demand. The same effect also improves the firm's

position in a wage bargaining process, since the firm can expect to fill a vacancy much quicker if a worker was laid off. Therefore, improved matching also leads to a reduction in wages.

ALMP programmes are also expected to have negative effects on the matching process, i.e. so-called locking-in effects. If a participation in an ALMP programme is associated with full time engagement, there might be insufficient time for actively searching for a regular job. In this case, the search effectiveness of participants is lower than the search effectiveness of the unemployed (Holmlund and Linden, 1993). Since this locking-in effect vanishes the moment the programme expires, do the positive effects on the search effectiveness persist after the participation has ended?

A second potential effect of ALMP is reduced welfare losses for the unemployed. If an ALMP programme increases re-employment probability or if the compensation level is higher than the unemployment benefits, the ALMP programme increases the expected welfare of the unemployed. This is caused by the fact that an unemployed person faces a positive probability of being placed in a programme with a consequent rise in the expected income. In the context of a wage bargaining process this is the same as an increase in fallback income, i.e. the income that is obtained if the bargaining fails and the worker becomes unemployed (Layard et al., 1991). The rise in the fallback income leads to a higher outcome for wages, since the position of the workers in the bargaining process is improved. This effect of ALMP on wage pressure is not avoidable, since every improvement of the situation of the unemployed is connected with a reduction in the welfare losses.

There is also a *competition effect*, in that ALMP (especially training programmes) are expected to improve the skills of participants and so make them more competitive. This means that not only is there improved competition between the unemployed, but also improved competition between the employed and the unemployed. Additionally, ALMP can affect competition if it stimulates participants to search more actively (i.e. to counteract the discouraged worker effect) or if it helps to increase labour-force participation. In both cases there is a rise in the effective labour supply, which leads to a reduction in wages.

⁽²⁴⁾ This is equivalent to an inward shift of the Beveridge Curve, i.e. a reduction of the unemployment rate for a given vacancy rate.

Finally, there are productivity effects. ALMP programmes that improve the skills of participants or serve as a substitute for work experience can be expected to improve or to maintain the productivity of participants.

The productivity effect refers to firms being able to produce more (or better quality) for given costs, particularly wages. Therefore, productivity is thought to be the marginal product of labour, i.e. the output of an additional hour of work. In an artificial model economy these measures would be equal to the wage rate. However, as most European economies face enormous wage rigidities, i.e. wages are not perfectly correlated with productivity. Considering a conventional labour-demand condition, a rise in productivity would lead to an increase in employment for a given wage rate. Calmfors (1994) notes that the rise in productivity is not self-evident, because there is also an opportunity to produce the same output with fewer, but more efficient, workers. Additionally, Calmfors et al. (2002) note that the rise in productivity of the participants may also have a wage-raising effect through a rise in the reservation wage of the participants.

Such a theoretical analysis should serve as a guideline for an empirical analysis of the impacts of ALMP. The impacts on matching efficiency, i.e. the estimation of aggregated matching functions, are the most frequent types of macroeconomic evaluations. Other types of empirical models are the reduced form relationships, in order to estimate the overall effect, i.e. the effect through the different channels, on employment or unemployment. Although such a model does not provide a differentiated picture of the effects of ALMP, it is able to quantify the net effect on the whole economy.

Another concern regarding macroeconomic evaluations is a serious simultaneity problem. In general, deciding how much money is spent on ALMP directly relates to the situation in the labour market. Spending on ALMP should, therefore, be determined by a policy reaction function. As a result, ALMP activity does not only determine the dependent variable in a macroeconomic evaluation; the dependent variable also determines ALMP activity. This classical simultaneity problem does not allow for identification of the parameters of ALMP measures. To solve this problem, instrumental variable (IV) estimators should be used for the estimation. Therefore, the main problem is to find valid and good instruments. The validity of

the instruments refers to the requirement that the instrument should not be determined by the dependent variable. The requirement for a good instrument is that the set instruments should be able to explain significantly the ALMP activity. From a practical point of view, the problem of finding an appropriate set of instruments is clearly a problem of data availability. Therefore, in most cases where not enough data is available, a good strategy is to use lagged values for the ALMP measures as instruments.

2.4. Cost-benefit analysis (CBA)

The previous sections introduced a number of econometric methods which can be used to assess the impacts of social programmes. Microeconomic evaluation studies are able to disclose whether social programmes have an impact on the individual. In addition, macroeconomic studies can cope with external effects that have to be taken into account. They are a necessary first step in assessing the value and merits of social programmes. If, for example, a microeconomic evaluation reveals that a certain programme has no impact at individual level, it hardly makes sense to complement it with a CBA and to assess its financial effectiveness.

Conducting a CBA widens the perspective of an impact analysis. A CBA is a method that provides a consistent, explicit and transparent procedure to evaluate public projects in terms of their consequences, i.e. in terms of their costs and benefits. The aim of a CBA is an efficient allocation of scarce resources, allowing it to serve as a normative tool in social decision-making. It is similar to a commercial profitability calculation conducted by private establishments. There are, however, also differences in that a CBA considers additional aspects, such as equity and distributional aims, and takes into account all costs and benefits to society as a whole.

The growth of public spending for social programmes has increased interest in a systematic evaluation in terms of costs and benefits. Originally, CBA was applied to technical projects, such as water resources, or engineering projects such as highways. It was not easy to transfer these techniques to social programmes because such programmes often involve impacts which

are hard to measure using market prices or even hard to assess. Additionally, social programmes often comprise distributional and equity aims which make some value judgements inevitable.

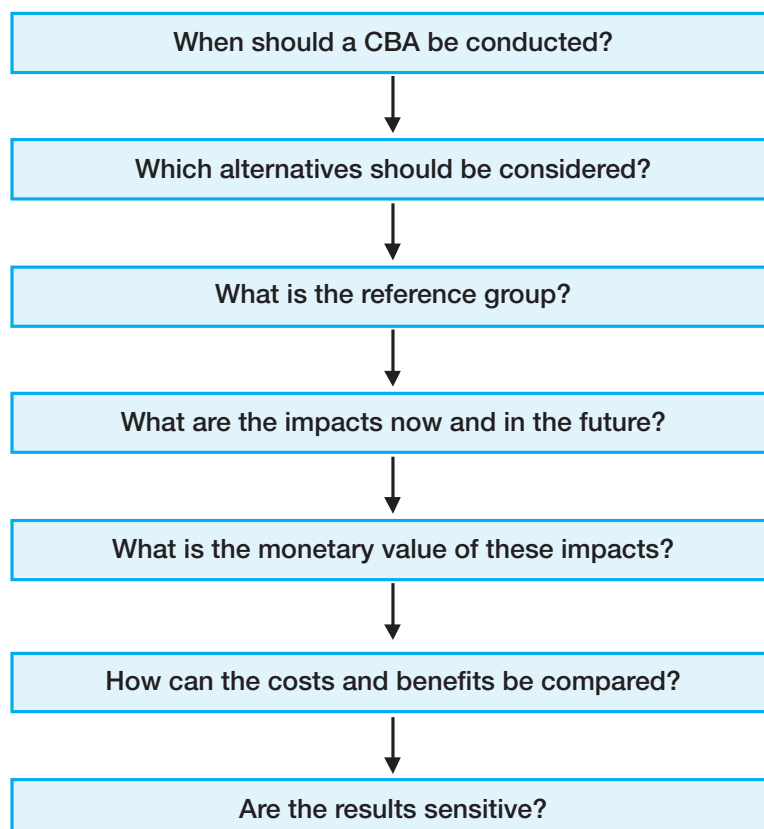
In the meantime, nearly all industrialised countries require such analysis for major social programmes (e.g. Boardman et al., 2001). The Unfunded Mandates Reform Act of 1995 requires that US agencies have to prepare an *ex ante* CBA for any regulation that may cost more than USD 100 million in any year. In Germany, *ex ante* CBA were established with the 1969 reform of the budget law which requires that CBAs have to be conducted for measures of considerable financial impact (Paragraph 7(2) *Bundeshaushaltsordnung*). *Ex post*, CBA are not explicitly regulated by law. Mandatory requirements for conducting *ex post* CBA do not exist in the US nor in Germany. However, in most cases, after programmes have been conducted officials find themselves having to justify these programmes. Thus, the intention to conduct similar programmes in the future depends on an *ex post* evaluation of the programmes.

Although CBA is a useful tool for assessing the efficiency of social programmes, it must be noted

that its role in practice is rather limited, especially in the field of ALMP (e.g. Delander and Niklasson, 1996). One reason for this is the problems and difficulties inherent in the method. We will address them later. Another point is that the use and acceptance of CBA depends heavily on the institutional and political setting in which it operates. A major feature of CBA is its thinking in terms of alternatives. If, however, the political situation prohibits some alternatives which are potentially superior to the actual situation, then a CBA is also limited and restricted in its results. CBA can play a constructive role if politics bear the burden of providing a rationale for any governmental intervention.

CBA can be defined as a systematic, explicit and transparent method to assess the net present value (NPV) of all benefits less all costs, valued by a single monetary measure, which resulted from a certain social programme. In this sense CBA is a method that quantifies, in monetary terms, the consequences of political decisions. This general definition calls for specification and clarification. The following aspects have to be fixed which also form the different stages of a CBA (see also Figure 4).

Figure 4: Different stages of a CBA



2.4.1. Time of a CBA

In deciding when to conduct a CBA, one has to choose between *ex ante*, *ex post* and *in-medias-res*, see Boardman et al. (2001), who also introduced a fourth category, namely one which compares an *ex ante* CBA with an *ex post* CBA. An *ex ante* CBA is conducted before the project is actually implemented. Its major advantage is that there is still time to change certain aspects of the programme. The result of an *ex ante* CBA indicates whether a certain programme should be implemented or not. If more than one programme has been evaluated, it can help in deciding which one to choose. Its major drawback is the fact that most of the benefits or costs will arise in the future and thus have to be estimated with uncertainty. Although plagued with these uncertainties, *ex ante* CBA are useful for decision-making purposes because it is still possible to make different use of the resources.

In contrast, *ex post* CBA are conducted when the programme has already been implemented and all costs are irreversibly 'sunk'. They provide more accurate and detailed information about a social programme since they do not rely on estimates and can also be used to learn more about similar programmes which are still to be implemented. It should be mentioned, however, that they suffer from the same problems as *ex ante* CBA when looking at the medium/long-term benefits of social programmes, which again have to be predicted. *In-medias-res* CBA are conducted during the life course of a social programme. They share some advantages and disadvantages of *ex ante* and some of *ex post* CBA. If, for example, there are only low sunk costs, *in-medias-res* CBA can be used to shift resources to other more desirable projects. If the sunk costs are quite high, they only might, as was the case for *ex post* CBA, give a concluding assessment of the project.

2.4.2. Selection of the alternatives

Selection of the alternatives concerns the identification of the project options to be evaluated. Every social programme aims at one specific target, such as improving the labour-market prospects of disabled persons (e.g. Delander and Niklasson, 1996). Alternative policy instruments to achieve this aim could be for example wage subsidies or increasing public employment opportunities. Although research could make

some proposals on potential alternatives, the available policy alternatives must be given by the policy-maker and are beyond the function of the analyst conducting a CBA.

This is important when interpreting the results of a CBA. If a CBA identifies one alternative as the optimal one, i.e. one with the highest present value of net-benefits, this only refers to the alternatives under consideration. There could be other alternatives with an even higher benefit which were not considered in the CBA.

2.4.3. Defining the reference group

Defining the reference group means deciding whose costs and benefits count, i.e. whose interests should be considered. Different perspectives are possible. One perspective is to differentiate between the target group, the non-target group and society as a whole. Additionally, one can include financiers as a separate group. A further alternative is the distinction between a local and global perspective. The choice of a certain perspective is given by the institutions who have ordered the analysis and is, therefore, not the function of the analyst. The perspective heavily influences which impacts have to be considered.

A profitability analysis in private establishments can be regarded as a CBA where only the private benefits, i.e. revenues, and private costs are considered. The net benefit is in this case identical to the firm's profits. A social CBA on the other hand, i.e. one which takes the perspective of the society as a whole, extends those private costs and benefits by including all impacts of the programmes whether they are private or social, tangible or intangible, direct or indirect.

The target group consists of those individuals who are eligible for the programme under consideration. If we consider a labour-market programme aimed at improving the labour-market prospects of disabled persons, the benefits for the target group consist of the effects of these programmes on the employment probability and wages. The costs, which have to be covered by this group, consist of earnings and transfers from which the participants have to abstain during the participation, i.e. opportunity costs of foregone income.

If we look at the group of non-participants, additional benefits and costs arise. Benefits which have to be taken into account in this perspective would be the additional output

attributable to the participants after they have finished the programme, the increasing tax payments, and more intangible impacts such as reduced delinquency. Costs which arise for this group are the expenditures for operating the programmes and the opportunity costs for the participants during participation. Aggregating these two views provides the social perspective which yields the net-benefit effect of the programme. This perspective yields the most complete, exhaustive and most complex view, since it requires the consideration of all possible impacts of a programme.

2.4.4. Enumerate and forecast all impacts

Probably the most crucial and important step in conducting a CBA consists of a complete enumeration of all impacts of a programme as costs and benefits, with a forecast of these impacts over the lifetime of the project. A condition for the identification of the impacts of a programme is the existence of a model which gives us the cause-and-effect relationship between the programme under consideration and the costs and benefits perceived by the target group. For some impacts this relationship is obvious, for example there is no doubt that measures for the disabled will influence their labour-market prospects or that the construction of a highway will reduce travel costs. The effects social programmes might have on more intangible factors, for example decreasing delinquency or more equity, are much harder to assess.

Impacts can be classified into real and monetary, direct and indirect, tangible or intangible, final or intermediate and, finally, internal or external effects. Real impacts mean the final impact on the social welfare, i.e. the final utility gain or loss of a programme, whereas monetary impacts only change relative prices while having no real welfare effects. Direct impacts refer to the intrinsic project goal while indirect effects cover those not intended by the programme. Examples of indirect effects in the context of labour-market programmes include the displacement effect or dead-weight losses. Final impacts occur at the level of the consumer while intermediate impacts occur at the producer level. The distinction between internal and external impacts is again closely related to the perspective of the CBA. Internal impacts are those which emerge within a pre-specified target group and which could also

be made up of external effects. These could occur, for example, if a programme which should bring the long-term unemployed back into work not only increases their employment prospects but also has other positive externalities such as crime reduction. Spillover effects and externalities are also responsible for some effects outside this group or area. Popular examples for external effects are air and noise pollution but could also contain positive externalities like those mentioned before.

The impacts to be considered depends heavily on the perspective chosen. In the example of the labour-market programme for the long-term unemployed, the target group benefits could be increasing employment probability after finishing the programme, therefore increasing income and life quality, reduced alcohol and drug abuse and reduced criminal activity (see for a practical application Long et al., 1981). Costs for the target group include forgone income while in the programme. In respect of society as a whole, additional external benefits, such as avoided costs for alternative services and additional costs, such as programme expenditure, have to be taken into account.

Other aims and objectives, which arise at the level of the society as a whole and which differentiate a CBA from private profitability calculations, are distributional and equity effects. The present value of a social programme only contains information on whether the benefits exceed the costs, i.e. it answers the question of whether the programme is efficient. Another important issue which has to be considered, however, is how these costs and benefits are distributed within society.

At this point a fundamental problem arises. In most cases, evaluating distributional and equity goals proves very difficult. Every publicly financed measure, for example in the area of ALMP, implies redistribution in that it transfers income from the non-participants to the participants of the programmes. Usually one Euro taken from non-participants is bestowed the same weight as one Euro given to the participants. One could, however, argue in this context that the marginal utility of one Euro is higher for the non-participants, who are assumed to have a lower income, than for the participants and that therefore the weights attached to the participants should be

higher. But since there are no objective and transparent methods to measure these weights accurately, they must be set within a political context and are therefore arbitrary and vulnerable to criticism.

2.4.5. Measuring and aggregating impacts

Having defined the impacts which have to be taken into account, the next step is to measure them in monetary terms. Measuring the impacts, i.e. 'monetising' them, is the heart of a CBA and also, in most cases, the most difficult step. For some impacts, for example the direct tangible effects of a programme, one can rely on market prices, whereas other impacts, such as saved lives, quality of life or equity are much harder to assess.

The simplest proceeding is an accounting of the monetary flows caused by a certain programme, for example expenditure on salaries or saved payments for social welfare. Another straightforward possibility, which is especially applicable in the case of tangible effects, is the direct observation of market prices. Although it is useful to start with market prices, very often it is necessary to adjust them in order to include external effects. If, for example, the social costs of labour input into a social programme should be evaluated, and if there is a large amount of involuntary unemployment, wages may have to be adjusted downwards in order to account for this idle input factor. This adjustment ensures that market prices correspond to the real net impact on welfare, i.e. to their shadow prices. Only in the case where there are no market distortions, i.e. when there is perfect competition and no external effects, will market prices correspond to their shadow prices. Market prices can also be used to assess the value of more indirect effects, such as declining delinquency. One possibility in this context would be to observe how prices for real estate evolve and to use the increase in these prices as an approximation for the utility of declining delinquency.

For intangible goods, like the value of a life saved or the increasing contentment of people who participate in a labour-market programme and afterwards find a job, referring to market prices is not feasible. In this case other methods have to be applied. One way is to directly ask for the amount of money individuals are willing to pay for a certain measure, for example an

improvement of the air quality or the like. To this end, a number of questions have been developed. By interpreting the results one has, however, to be aware of the hypothetical feature of these questions and therefore their limited applicability. Another way to measure intangible goods is the observation of political preferences. By observing, for example, that in the past a life-saving programme had been established that cost EUR 100 000 and was able to save 10 lives, one could infer that the value of a life amounts to EUR 10 000. Again such indirect inferences have to be made with care.

When a social perspective is chosen, one additional problem that arises is the aggregation of individual costs and benefits. We have addressed this issue already and also mentioned the problems and difficulties arising in this context. Aggregating the individual costs and benefits presupposes attaching weights to each individual. Due to the absence of other convincing weighting schemes, equal weights are usually used. One justification for this simplification is the following: when the net benefit of a social programme is greater than zero, the sum of all individual benefits exceeds the sum of all individual costs. In other words, the size of the 'common cake' increases (e.g. Delander and Niklasson, 1996). If this is the case, there is room for redistribution in the sense that the losers of the programme are compensated by the winners. We should note, however, that this indemnification view is only theoretical and one has to consider that with such redistribution policies, additional costs might arise which then will influence the net benefit of the programme.

If it turns out that the problem of measuring the utilities of a social programme cannot be resolved satisfactorily, then at least a cost-effectiveness analysis (CEA) can be conducted. Cost-effectiveness analysis evaluates programmes by comparing the costs associated with these programmes with a single quantified but not monetised effectiveness measure. It may be that in evaluating a health care programme the lives saved by alternative programmes have to be assessed. Conducting a CBA would require the valuation of the lives saved using monetary units. Instead, the analyst could also assess the different programmes by relating their costs to the lives saved, i.e. by constructing cost-efficiency

ratios. Another example of a cost-effectiveness analysis will be given later on.

2.4.6. Comparing costs and benefits

Once all costs and benefits are itemised and measured, another problem which has to be resolved stems from the fact that costs and benefits do not occur at one point in time but rather follow a dynamic path. In order to compare costs and benefits which occur in the future with those which occur now, it is necessary to define a discount rate r . Once a discount rate is found, the flows of costs C_t and benefits B_t can be discounted to give the net present value NPV of a social project according to:

$$NPV = \sum_t \frac{B_t - C_t}{1+r} \quad (32)$$

A simple decision rule is to implement a project if it has a net present value which is larger than zero and to reject it otherwise. If more than one project is evaluated, the one with the largest NPV should be implemented.

The choice of an appropriate discount rate, however, is contentious. One approach could be to use some market interest rate as an approximation to the social discount rate. The market interest rate, however, does not correspond to the social rate. This is because capital markets are not perfect and thus the time preferences of individuals do not correspond to market interest rates. Other points responsible for this departure are, for example, uncertainty about future inflation and the existence of tax effects which will be reflected in market rates.

In most programmes, costs occur during implementation while the benefits are spread over the future. Thus small changes in the discount rate might have a great impact on the net present value of the project. The choice of an appropriate discount rate is crucial in conducting a CBA and therefore an eligible candidate for a sensitivity analysis.

An alternative decision rule, which does not rely on a certain discount rate, is to determine the internal rate of return of the social programme. The internal rate of return r of a social programme is

equal to the discount rate for which the net present value of the social programme becomes zero:

$$NPV = \sum_t \frac{B_t - C_t}{1+r} = 0. \quad (33)$$

When more than one project is evaluated, the one with the highest internal rate of return is chosen. When only one project is considered, again a reference value is needed. If the internal rate of return is above this reference value, the project should be realised. If not, it should be rejected.

This method, however, has some drawbacks. If the flow of net-benefits alters between positive and negative values during the lifetime of the project, more than one value for the internal rate of return might fulfil the above equation. Additionally, decisions which rely on this rule might differ from decisions derived by using the net present value of the projects.

2.4.7. Conducting sensitivity tests

While presenting the necessary steps for conducting a CBA, there are various caveats and drawbacks which might heavily influence the results and recommendations. CBA relies on a number of assumptions and uncertain predictions, for example, with regard to the potential impacts, the way to measure them and at which interest rate to discount them. All these topics are eligible for a sensitivity analysis which examines how sensitive the NPV results are to different assumptions about those key parameters. Careful scenarios, which vary the most important assumptions, are one way to protect the results against indicated doubts.

In order to illustrate the above, one hypothetical and one real case study in the field of ALMP will be used. Let us assume that in a hypothetical situation an economy tries to fight its rising unemployment by running different measures of ALMP⁽²⁵⁾. More specifically, four different instruments will be considered: LMT, youth measures, subsidised employment and measures for the disabled. The following table contains hypothetical figures for expenditure on these programmes and for the number of participants in the different measures. Total expenditure amounts to EUR 33 million, with participation by 15 753 individuals.

⁽²⁵⁾ Another assumption is that we are only interested in a reduction of overall unemployment. Other programme objectives are not included in this simple example.

Table 4: **Hypothetical example before reallocation**

Programme	Expenditures (Million EUR)	Participants	Per capita expenditures (EUR)	Treatment effect (%)	Employed participants
LMT	5	2 772	1 804	80	2 218
Youth measures	10	5 188	1 928	20	1 038
Subsidised employment	8	5 273	1 517	5	264
Measures for disabled	10	2 520	3 968	5	126
Total	33	15 753	2 095	23	3 645

Let us now assume that a microeconomic evaluation study has been conducted and that this study revealed that the various measures had different impacts on individual employment probability after finishing the various programmes. LMT measures were most successful in fighting unemployment with a treatment effect of 80 %. This means that participants in LMT measures had an 80 % chance of finding a new job after finishing the programme compared with an appropriate control group of non-participants. The treatment effects of the other programmes can be interpreted in a similar way.

Our hypothetical example indicates that the EUR 33 million spent enable 23 % of the 15 753 participants to find new employment. The cost

efficiency ratio before an efficient reallocation is thus 23 %. The findings of the microeconomic analysis can now be used to re-allocate expenditure more efficiently. Let us consider that, on the basis of these findings, the government has decided to expand the measures of LMT and youth measures by 100 % and 50 % respectively, and at the same time to decrease expenditure on the other two labour-market programmes by 75 % respectively. This reallocation could be achieved by a more generous or a more restrictive interpretation of the regulatory laws. The following table contains the expenditure and the number of employed participants after such a re-allocation ⁽²⁶⁾.

Table 5: **Hypothetical example after reallocation**

Programme	Expenditures (Million EUR)	Participants	Per capita expenditures (EUR)	Treatment effect (%)	Employed participants
LMT	10	5 544	1 804	80	4 435
Youth measures	15	7 782	1 928	20	1 556
Subsidised employment	2	1 318	1 517	5	66
Measures for disabled	2.5	630	3 968	5	32
Total	29.5	15 274	1 931	40	6 089

⁽²⁶⁾ This allocation is not simply a replacement of more expensive with cheaper measures but a reallocation between less and more efficient ones. Even though youth measures, for example, are more expensive than subsidising employment (both in the total amount and per capita spending), the first measure is the more efficient one and thus should be expanded.

Even with a smaller amount of money, more participants were put back into the workforce and the average treatment effect of all measures has nearly doubled. Since an econometric evaluation also delivers other valuable information, for example for which subgroups of participants which programmes are especially successful, this gives a first impression of the value and the usefulness of econometric evaluation studies.

Another real example of a CBA (Long et al., 1981 and Delander and Niklasson, 1996) aimed to assess the effects of a US federal social programme for economically disadvantaged youths, the Job Corps programme. This programme is a comprehensive set of services

for disadvantaged youths, such as vocational skill training, basic education and health care. The aim of the programme was to increase the employability of the participants.

In order to take account of distributional effects, three different perspectives have been considered: the group of participants, the group of non-participants and an aggregated view of society as a whole. Focusing on the perspective of society as a whole addresses the issue of efficiency while hints about the distributional consequences of the programme can be obtained by looking at the two groups separately. A wide range of potential benefits and costs have been taken into account as summarised in the following table:

Table 6: **Costs and benefits of the Job Corps programme**

Benefits	Examples
Output produced by Corps members	In-programme output, increased post-programme output, increased tax payments, increased utility due to preferences for work over welfare.
Reduced dependence on transfer programs	Reduced transfer payments, reduced administrative costs.
Reduced criminal activity	Reduced criminal justices system costs, reduced personal injury and property damage.
Reduced drug/alcohol abuse	Reduced treatment costs, increased utility from reduced drug/alcohol dependence.
Reduced utilisation of alternative services	Reduced costs of other programmes than the Job Corps.
Other benefits	Increased utility from redistribution, increased utility from improved well-being of Corps member.
Costs	Examples
Programme operating expenditures	Centre operating expenditures, excluding transfers to Corps members, central administrative costs.
Opportunity costs of Corps-member labour	Forgone output, forgone tax payments.

The effects of the programme were estimated using data from a survey conducted among the group of participants and a comparison group who were never enrolled in this programme. Since the programme was not constructed as a social experiment, multiple regression techniques, which account for both observable and unobservable effects, had to be employed to estimate the treatment effect of the programmes.

The other costs and benefits were valued so that they reflect the resources saved, consumed or produced as a result of the Job Corp

programme. The benefits arise primarily from two factors. First, Corps members made less use of other social programmes and committed fewer crimes. Second, corps members improved their long-term employment prospects and increased their contribution to the gross national product (GNP) and tax payments. The effects of these changes were valued by multiplying the estimated change in the behaviour of the participants due to the programme by estimated dollar values.

The contribution of the participants to GNP, for example, was estimated by using the difference

between the earnings of the Job Corps members and the non-participants. The implicit assumption in this procedure is that labour markets are competitive, so that the wages are equal to the marginal contribution to the GNP, and that there are no displacement effects due to the programme. Using this method, the authors estimated that the total discounted value of the increased output over the first two years after the end of the programme is USD 925 per participant.

The benefits stemming from reduced delinquency were estimated by multiplying the estimated changes in arrests due to the programme by the shadow prices, which are equal to the cost savings per avoided arrest. The shadow prices were calculated by considering the costs caused by arrested persons, for example police custody, arraignment, detention, etc. Costs of the project

mainly comprised operating expenditures for the programme centres and central administration costs. They represent costs to non-participants. Other categories which do not appear in the financial accounts are forgone income of the participants and forgone tax payments.

Making a set of assumptions, for example setting the discount rate to 5 %, assuming the effects of the programme to diminish at a rate of 50 % every 5 years, the present value of the net benefits can be estimated. The results suggest that the programme yields a net benefit to society and also to the target group. The group of non-participants experiences a slightly negative value from the programme. The programme was estimated to represent a socially efficient use of the resources and additionally entails redistribution from the group of non-participants to the group of participants.

3. ALMP in the EU and empirical findings

3.1. Recent developments in ALMP in the EU

ALMP have been seen as one way to fight the rising unemployment rates and the disequilibrium (skill shortage, etc.) in the labour markets which have arisen in most European countries since the early 1970s. Table A.1 in the annex shows standardised unemployment rates in the EU from 1985 to 2000. In 1985 the EU average was 10.8%. After declining in the following years, it came back to this value in 1994. In subsequent years unemployment eased, bringing the rate down to 6.9% in the year 2000 ⁽²⁷⁾. The average reduction from 1985 to 2000 was 36%. It is quite interesting to compare this with the picture in the different countries. Sweden and Finland had to deal with extraordinary increases in unemployment during this period. The rate rose by 108% in Sweden (from 2.8% to 5.9%) and 94% in Finland (from 5.0% to 9.7%). In France there was a slight increase from 8.3% to 9.3% (+12%), whereas in all the other countries unemployment has fallen since 1985. The biggest decreases can be found in Ireland (-76%, from 17.7% in 1985 to 4.2% in 2000) and the Netherlands (-82%, from 13.1% to 2.8%). The remaining countries had decreases between 15% (Germany) and 52% (Portugal).

Whether these reductions are due to ALMP or other factors has to be determined by evaluation studies, which will be reviewed in the next chapter. We have pointed out already, that the importance of ALMP has been growing enormously from 1985 to 2000. The ratio between ALMP and PLMP has risen from 0.44 in 1985 to 0.68 in 2000 (EU average). Even though there is an EU-wide increase in ALMP this is not true for all of the countries. Therefore we need to take a closer look at the evolution in specific countries. Tables A.2 and A.3 in the annex show the spending on ALMP and PLMP for the years 1985 to 2000. If we look at Table A.3 we see that spending on PLMP peaked in the mid-1990s in line with the peak in unemploy-

ment. The obvious reason is that unemployment benefits are entitlement programmes, i.e. rising unemployment automatically increases public spending on passive income support. Active labour-market programmes, on the other hand, are discretionary in nature and therefore more easily disposed with in a situation of tight budgets. Despite that, ALMP spending reached its highest level in the mid-1990s; a clear indication that ALMP were seen as a suitable measure against the bad labour-market situation.

Figures 5 and 6 compare spending on ALMP and PLMP in the years 1985 and 2000. The bisecting line in the figures indicates a balanced relationship between both measures, whereby countries in the left/right half direct more money into passive/active measures respectively. It is interesting to note that the differences between the countries were more pronounced in 1985. Italy and Sweden spent more money on active measures, spending on PLMP by Belgium, Denmark, Ireland and the Netherlands exceeded spending on ALMP by more than two percentage points of GDP. In 2000 the situation is much more balanced and there is a tendency to an equal spending on both measures. Italy still spends more money on active measures, Greece and Sweden spend equally for ALMP and PLMP. The other countries all spend more money on PLMP but they are moving closer to an equal spending. Denmark still spends most money on PLMP (3% of the GDP), but the difference between PLMP and ALMP was only 1.46 percentage points in 2000, compared to 2.7 percentage points in 1985.

Having discussed the importance of ALMP in contrast to PLMP in general, we look now at the importance of one special programme, LMT. Table A.4 in the Annex shows the spending on LMT as a percentage of GDP for the years 1985 to 2000. We see that Denmark (0.72%) and Sweden (0.64%) direct the highest amounts of their GDP to LMT. Table 7 demonstrates the relative importance of LMT, by showing the spending on LMT as a

⁽²⁷⁾ Unfortunately the OECD does not provide standardised unemployment rates for Austria and Greece until 1993.

Figure 5: Spending on ALMP and PLMP in the EU, 1985

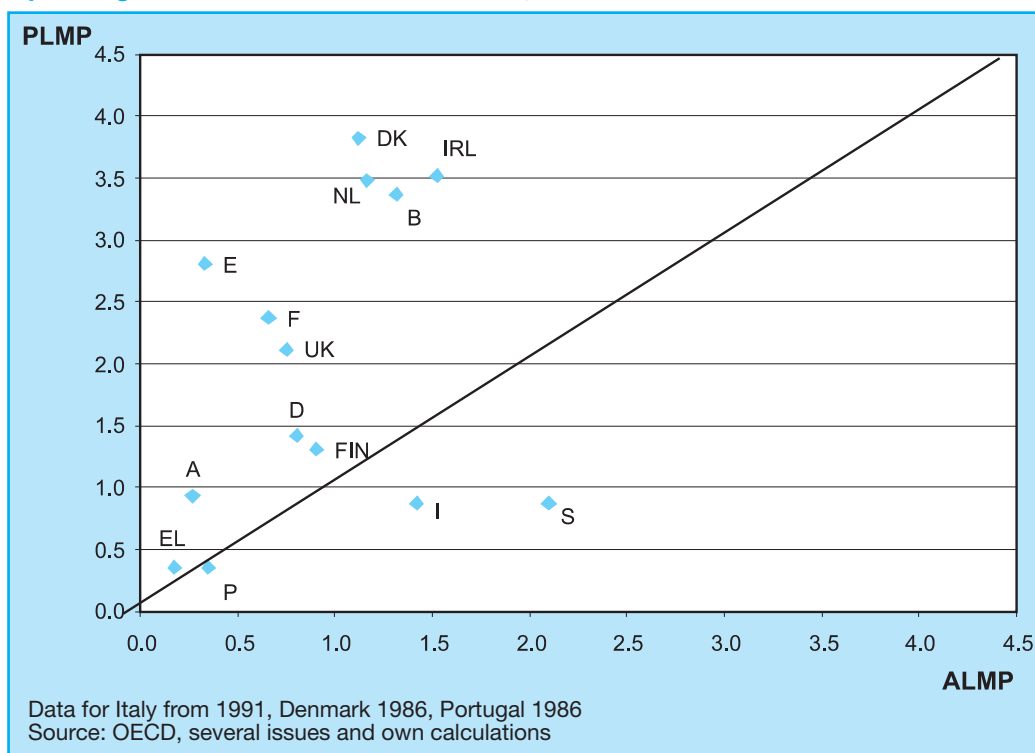
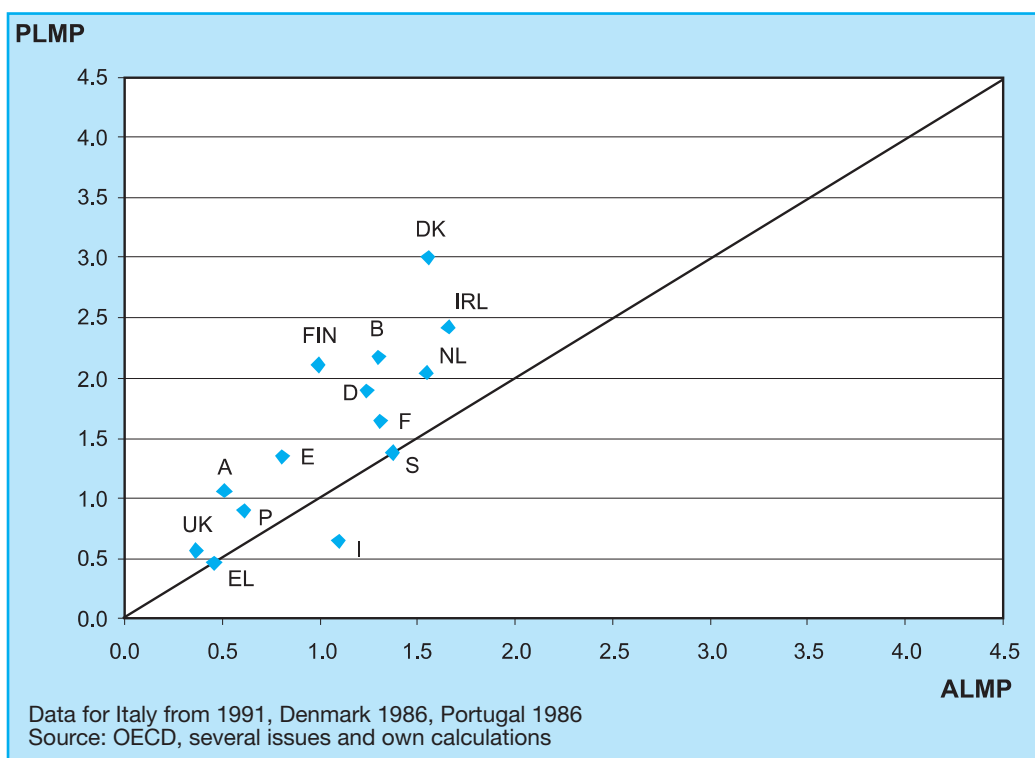


Figure 6: Spending on ALMP and PLMP in the EU, 2000



percentage of the total spending on ALMP. Denmark directs 43 % of its spending on ALMP to LMT, in contrast to Italy's 5 %. The EU average spend on LMT is 0.3 % of GDP and this corre-

sponds to 26 % of the total spending on ALMP. The justification for this expenditure has to be seen from the empirical evaluation of these programme studies and this will be done in the next chapter.

Table 7: **Expenditure on LMT as a percentage of total spending on ALMP**

(%)

	1985	1986	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	Average
Austria	30	34	31	33	29	32	31	33	38	39	33	36	33	33
Belgium	15	15	17	19	19	22	21	20	19	21	18	18	18	19
Denmark	..	38	24	28	27	27	40	52	60	56	58	56	54	43
Finland	29	29	25	24	26	28	28	28	33	35	32	33	30	29
France	39	36	41	38	36	35	32	29	26	25	23	21	19	31
Germany	25	26	37	35	38	35	30	28	31	29	27	26	27	30
Greece	12	14	44	52	32	26	18	29	19	18	62	..	..	30
Ireland	41	36	32	19	..	..	15	13	13	..	..	..	..	24
Italy	..	..	..	1	1	1	1	1	1	12	13	11	..	5
Netherlands	11	16	21	22	25	23	27	26	23	24	20	21	19	21
Portugal	..	51	20	24	31	30	30	29	33	35	38	36	25	32
Spain	6	11	19	22	18	23	36	38	50	31	24	17	19	24
Sweden	24	25	31	41	35	26	26	24		20	23	27	23	27
UK	9	8	34	26	27	26	25	22	21	19	18	15	214	20
Means	17	24	28	27	27	25	26	27	29	29	29	28	27	26
..=Missing values Source: OECD Employment Outlook, 2002.														

3.2. Microeconomic evaluation studies

To see how successful labour training programmes have been in the recent years, we have reviewed 45 studies from 10 European countries. The results from microeconomic evaluation studies in Germany and Europe will be presented before those of macroeconomic studies. The results of the studies, the outcome variables used, the programmes evaluated and the methods applied are also summarised in Table 8 at the end of this chapter.

3.2.1. Germany

All evaluation studies of vocational training for West Germany are based on the German Socioeconomic Panel (GSOEP)-West. With respect to the design of the programmes, measures on-the-job and off-the-job are considered as well as measures with and without income maintenance. The applied evaluation methods include discrete hazard rate models, matching, instrumental-variable estimation and simultaneous dynamic models. Besides unemployment duration and the re-employment probability, hourly wages and employment stability are considered as outcome variables. Hujer et al.

(1998, 1999a) find that participation in vocational training has significant effect in reducing unemployment duration. In further studies, Hujer et al. (1999b), and Hujer and Wellner (2000b) discover positive effects only for short courses (< 6 months), whereas long courses do not have (significant) positive effects. Hujer et al. (1999c) found that on-the-job training has no significant effect on unemployment duration, whereas off-the-job training reduces it in the short-term and has no significant effects in the long-term. This finding corresponds to Pannenberg (1995, 1996), who observes that participation in off-the-job training increases re-employment probability in the short term. In contrast to this positive finding, the following studies produce rather negative results. Prey (1997) examines vocational training with income maintenance and finds negative (no) effects for men (women) on the employment probability. In a further study, Prey (1999) finds negative (no) effects on employment probability for measures with (without) income maintenance and no effects on the wages. Staat (1997) studies public sector training with income maintenance. His results indicate positive effects on the search duration only for subgroups, no significant effects on the employment probability but positive effects on wages.

The studies for East Germany are either based on the East German Labour Market Monitor (EGLMM), the GSOEP-East or the Labour Market Monitor Saxony-Anhalt, a panel based on the population of the state of Saxony-Anhalt. As with the studies for West Germany, training is analysed on- and off-the-job, as well as with and without income maintenance. The period under consideration ranges from 1989 to 1998. There is examination of several outcome variables; (un)employment duration, (stable and unstable) employment (probabilities), job search, working time and wages. Fitzenberger and Prey (1997) find that training outside the firm shows considerable positive effects on employment probabilities, whereas training inside has negative effects. Hübler (1998) found that on-the-job training increases job security, whereas off-the-job training leads to higher earnings. But this is true only for privately financed training. Publicly financed training has only positive effects in the short-term. Fitzenberger and Prey (2000) question training supported by public income maintenance outside a firm, and discover positive effects for employment and earnings (but only a few significant in the long-term). Pannenberg (1995, 1996) ascertains positive effects on the re-employment probability and the income for vocational training on- and off-the-job. Lechner (1999b) examines enterprise-related continuous vocational training and finds positive income effects too, but no effects on unemployment probability. Kraus et al. (1998) observe, for a sub-period of their study, positive effects on (stable) employment for on- and off-the-job training. Hübler (1994) examines on-the-job qualifying measures and finds that training induces search activities and reduces the effective hours of work.

Many studies do not find any clear positive or negative effects. Bergemann et al. (2000) examine multiple participation in further training. They discover no positive effects for first training programmes, and the additional effects of a second participation are, on average, not different from zero. Hujer and Wellner (2000a) find no significant effects for public-sector sponsored vocational training on unemployment duration, but very weak hints that short courses seem to be more effective in reducing the duration. Lechner (1999c) investigates off-the-job training (publicly financed and enterprise related). He cannot establish robust positive effects on either

employment probability or the earnings. Staat (1997) examines public sector training with income maintenance and finds no effects either on the search duration, or on the employment stability or the level of hourly wages.

The results of the following studies tend to be negative. Fitzenberger and Prey (1998) examine the effects of training within and outside the firm and with and without public income maintenance on employment and wages. Mostly they get negative or no effects, differing with respect to the specification. Lechner (2000) finds no positive long-term effects on employment probabilities or earnings for public-sector-sponsored vocational training. He gets negative results regarding the risk of unemployment in the short-term.

3.2.2. Other European countries

This section presents an overview of microeconomic studies in Europe. Due to numerous microeconomic evaluations in Europe, the following section will focus only on recent studies that, in addition, deal with training programmes. Further overviews of European studies on the microeconomic evaluation of ALMP are given, for example, by Steiner and Hagen (2002) and Heckman et al. (1999).

3.2.2.1. Switzerland

Lalive et al. (2000) analysed vocational training and job creation programmes in Switzerland. They found that both programmes have negative effects on employment during participation. Furthermore, they found that these negative locking-in effects seem to dominate the positive effect on employment after the programme has expired. This problem seems to be particularly severe for job creation schemes for men, whereas for women it seems to be less important.

Another evaluation for Switzerland was performed by Gerfin and Lechner (2000). They also analysed different vocational training and job creation programmes, including training to improve basic skills, language and informatics abilities as well as further training. Their results indicate that the informatics courses and further training programmes have no effect on employment and that the basic skills and the language courses seem to have a negative effect. Furthermore, they found that unemployment duration prior to the programme participation seems to have an influence on the programme effect, i.e. programmes are

more effective for long-term unemployment (LTU). For the job creation programmes they found negative effects on employment, whereas positive effects were found for women.

3.2.2.2. *Austria*

Zweimüller and Winter-Ebmer (1996) analysed the effects of Austrian vocational training on employment stability. Their results indicate that training programmes significantly reduce unemployment risk for the participants.

Winter-Ebmer (2001) analysed the employment effects of job creation and training programmes by the *Österreichischen Stahlstiftung* on employment. These programmes are thought to counteract staff reduction within the steel industry, so only steelworkers are eligible for the participation. The results of Winter-Ebmer (2001) indicate that there are positive employment effects primarily for unemployed older than 27 years, whereas for younger people there are no effects.

3.2.2.3. *Belgium*

In a study for Belgium, Cockx et al. (1998) compared the effects of training programmes and subsidised employment programmes on employment duration. They analysed in-house and external training programmes. Whereas they found no effect from subsidised employment and external training programmes, they found a positive effect from in-house training programmes. As internal training programmes are an investment in firms' human capital, they should reduce the risk of being laid off.

In a subsequent study, Cockx and Bardoulat (2000) analysed the effect of vocational training programmes on exit from unemployment evaluating training programmes not organised by firms. They found a negative locking-in effect during the programme and a positive effect after the programme had expired. The positive effects subsequent to the programme, compensate for the locking-in effect and thus the overall effect was positive.

3.2.2.4. *France*

French vocational training and job creation programmes were analysed by Bonnal, Fougère and Sérandon (1997). In their analysis they distinguished between unemployed people with vocational education and those without, finding a positive effect from training programmes on exit

from unemployment for the unemployed without vocational education. They found no positive results from job creation programmes. Brodaty et al. (2001) obtained similar results in their study where they used the same dataset but different evaluation methods.

3.2.2.5. *Denmark*

Jensen et al. (1999) analysed a youth unemployment programme that was established in Denmark in 1996. Unemployed youths without vocational education are obliged to participate in a vocational training programme to remain entitled to unemployment benefits. Jensen et al. (1999) analysed the effects of this programme on entry into regular employment and into a regular vocational education. While they find no significant effect on entry into regular employment, the effects on entry into regular vocational education are positive.

3.2.2.6. *United Kingdom*

Firth et al. (1999) analysed the effects of Employment Training (a vocational training programme) and Employment Action (a job creation programme) on the separation rate into employment in the United Kingdom. Employment Training offered a significant positive effect whereas for Employment Action no significant effects could be found. The same results were also obtained by the prior study from Payne et al. (1996) that analysed the effects of both programmes on employment.

3.2.2.7. *Ireland*

For Ireland, O'Connell and McGinnity (1997) analysed the effects of classroom training, on-the-job training and employment subsidy programmes on employment. Their results indicate that both training measures significantly increase employment, whereas for the employment subsidy programme no significant effect could be found.

3.2.2.8. *Norway*

Aakvik, Heckman and Vytlačil (2000) analysed the effects of subsidised employment and training programmes on employment for women in Norway with long-term diseases. All programmes recorded a negative effect. The authors note proposed that the negative effect may have been caused by the selection of the participants. It seems that those placed into the programmes

were mostly unemployed women who had good employment opportunities anyway.

3.2.2.9. Sweden

Compared to other European countries, Sweden has a long tradition of ALMP. Consequently, the Swedish active policy measures were evaluated more frequently than other European countries. Presented here are only three of the latest microeconomic evaluations for Sweden. A complete overview can be found in Calmfors et al. (2002).

Johansson and Martinson (2000) compared traditional vocational programmes (organised by the labour administration) with a special training programme to qualify the unemployed for jobs in the information technology sector (Sweden Information Technology – SwIT). This training programme is carried out in cooperation with industry in order to ensure a sufficient supply of qualified workers. Analysing the effects of both programmes, Johansson and Martinson (2000) find that the Sweden Information Technology programme increases employment probability more than the traditional training programmes. The authors see this as evidence that a more focused contact with firms can increase the efficiency of training programmes.

Larsson (2000) analysed the effects of job creation and training programmes for young people. The study finds no evidence that there is a positive effect from either set of programmes on regular employment or on regular vocational education.

Sianesi (2001) analysed the effects of ALMP programmes without a differentiation between the different programme types, finding a positive impact on the registered unemployed. A serious problem seems to be the fact that participation serves as a requirement to renew entitlement for unemployment benefits. Finally Richardson and van den Berg (2001) analysed (in an empirical study) the impact of employment training in Sweden targeted at unemployed individuals as well as employed persons who are at risk of becoming unemployed. The results show highly positive effects, with a doubling of individual re-employment rates after completion of training. Regarding the time spent within the programme, the individual net effect on unemployment duration reduces to zero.

3.3. Macroeconometric evaluation studies

Macroeconometric evaluations are rare in Europe, i.e. for most European countries there are no studies. As a single time series at national level usually does not provide enough observations, most studies rely on pooled cross-section time series data. Primarily these are studies using regional data to evaluate ALMP for one specific country. Besides these studies there are also some using a cross-country dataset for the OECD countries (e.g. Jackman et al., 1990 or OECD, 1993). Since this data, and hence the studies, are not restricted to Europe, statements for Europe alone cannot be made. Furthermore, cross-country studies suffer from the problem that they are supposed to analyse heterogeneous policy measures, which is a major drawback of these studies.

3.3.1. Sweden

Sweden provides the most macroeconomic evaluations using regional data. An overview can again be found in Calmfors et al. (2002). A prominent study for Sweden was conducted by Calmfors and Skedinger (1995). They used a reduced form relationship to analyse the effects of job creation and training programmes on the total rate of job seekers (i.e. openly unemployed and programme participants relative to the labour force). It is worth noting that they have looked specifically for the simultaneity problem of ALMP. As instruments they used not only labour-market indicators but also political factors such as the proportion of seats in the parliament assigned to left-wing parties. Their results indicate that job creation schemes tend to crowd out regular employment and that the results for vocational training, although they are unstable, are more favourable than the results for job creation schemes.

3.3.2. Germany

In addition to the studies for Sweden, there are some studies for Germany that were conducted with regional data. Büttner and Prey (1998) evaluated training programmes and public sector job creation for West Germany. They find that job creation schemes reduce mismatch, whereas training programmes do not have any significant effects. Prey (1999) extends this work by addi-

tionally controlling for the regional age structure and recipients of social assistance and estimating separately for men and women. She finds that vocational training increases (decreases) the mismatch for women (men), whereas job creation scheme decreases the mismatch for men.

Pannenberg and Schwarze (1998) use the data from 35 local labour office districts to evaluate training programmes in East Germany. They find that the programmes have negative effects on regional wages. Schmid et al. (2000) estimate the effects of further training, retraining, public sector job creation and wage subsidies on long-term unemployment and exit from unemployment. They find that job creation schemes reduce only 'short' long-term unemployment, vocational training reduces long-term unemployment and wage subsidies help only the very long-term unemployed. Steiner et al. (1998) examine the effects of vocational training on the labour-market mismatch in East Germany. They observe only very small effects on the matching efficiency which disappear in the long term. Hagen and Steiner (2000) evaluate vocational training, job creation schemes and social assistance measures (SAM) for East and West Germany using the data from local labour office districts. The estimated net-effects are not very promising as all measures increase unemployment in West Germany. Only social assistance measures reduce unemployment slightly in East Germany, whereas job creation schemes and vocational training increase it too.

3.4. Vocational training programme success in previous years

Table 8 summarises empirical microeconomic studies for Germany and Europe. The first thing to note is that there seem to be more favourable effects of training programmes in general compared to other types of programmes. Since training programmes are one of the most important measures of ALMP in Europe they are often compared to other measures such as job creation schemes or subsidised employment programmes. Whereas we find positive effects for vocational

training programmes in most cases, the effects of other measures are positive only in one case. Furthermore we find that training programmes that are organised by firms seem to be more effective compared to publicly organised programmes. This may be reasoned by a closer contact with the firms which may increase the efficiency in adjusting the skills of the participants to the demands of the firm.

We find immense differences between East and West Germany. In West Germany the results for vocational training programmes look promising, especially evaluations with respect to the duration of (un)employment. Unfortunately this positive effect does not remain stable if the evaluation is based on other outcome variables. Studies using alternative outcome variables, such as re-employment probability or wages, show a very different picture. An interesting finding is the result from Hujer et al. (1999b), and Hujer and Wellner (2000b) where short (< 6 months) training programmes were found to be more effective than long ones. This may be caused by a negative selection effect, i.e. the unemployed that need comprehensive training to become eligible for a regular job are *a priori* disadvantaged in order to find a regular job. If such a negative selection effect is present it is clearly important to differentiate between different lengths of training programmes in an evaluation. Turning to East Germany we find a positive effect in most studies of vocational training, whereas there are hardly any studies that indicate a negative effect.

In evaluation it should be noted that the choice of the outcome variable seems to be an important issue. Outcome variables like earnings or wages may be questionable if we are evaluating European programmes. This is due to the fact that the welfare state and minimum wage regulations are responsible for distortions between the employment status and the earnings. Therefore, outcome variables that are directly associated with employment status, such as employment probability or the unemployment duration, are preferable for evaluations in Europe.

In this context it might be useful not only to regulate the evaluation by law but also to give some recommendations with respect to the outcome variable which should be used ⁽²⁸⁾.

⁽²⁸⁾ It should be noted that in Germany regulatory law as in the Social Code III, provides the mandatory use of outcome variables, for example re-employment.

Table 8: Summary of the empirical findings from microeconomic studies

Country/Author	Outcome variable	Programme	Result*	Applied method
Austria				
Winte-ebmer (2001)	Employment	Vocational training	(+)	Tobit / IV
Switzerland				
Lalive et al. (2000)	Inflows into employment	Vocational training Job creation schemes	(-) (-) men / (0) women	Duration analysis
Gerfin and Lechner (2000)	Employment rate	Training programmes - basic training programmes - language training - informatics courses - vocational training Job creation schemes	(-) (-) (0) (0) (-) men / (+) women	Matching
Belgium				
Cockx et al. (1998)	Employment duration	Training programmes: - in-firm - external Subsidised employment programmes	(+) (0) (0)	Duration analysis
Cockx and Bardoulat (2000)	Outflow rate from unemployment	Vocational training programmes	(+)	Minimum Chi Squares / IV-estimation
France				
Bonnal et al. (1997)	Outflow rate from unemployment	Vocational training Job creation schemes	(+) (0)	Duration analysis
Brodaty et al. (2001)	Outflow rate from unemployment	Vocational training Job creation schemes	(+) (0)	Duration analysis
Denmark				
Jensen et al. (1999)	Inflows into regular employment	Youth unemployment programme (vocational training)	(0)	Duration analysis
	Inflows into regular vocational education		(+)	
United Kingdom				
Firth et al. (1999)	Inflows into regular employment	Employment training Employment action	(+) (0)	Duration analysis
Payne et al. (1996)	Employment rate	Employment training Employment action	(+) (0)	Matching / selection
Ireland				
O'Connell and McGinnity (1997)	Employment rate	Classroom training On-the-job training Employment subsidies programmes	(+) (+) (0)	Probit

Norway				
Aakvik et al. (2000)	Employment	Training programmes Subsidised employment programmes	(-) (-)	Discrete choice
Sweden				
Johansson and Martinson (2000)	Employment rate	a) Vocational programmes b) Information technology training programmes	(+) b>a	Matching
Larsson (2000)	Employment	Vocational training Job creation schemes	(0) (0)	Matching
Sianesi (2001)	Unemployment	ALMP in general	(+)	Matching
Richardson and van den Berg (2001)	Employment rate	Employment training programme	(+)	Duration analysis
Germany (West)				
Hujer et al. (1998, 1999a)	Unemployment duration	Vocational training	(-)	Duration analysis
Hujer et al. (1999b), Hujer and Wellner (2000b)	Unemployment duration	Vocational training - short courses (< 6 months) - long courses	(-) (0)	Duration analysis
Hujer et al. (1999c)	Unemployment duration	Off-the-job training Off-the-job training	(0) (-)	Duration analysis
Pannenberg (1995, 1996)	Re-employment probability	Off-the-job training	(+)	Discrete hazard rate, logit and probit estimates
Prey (1997)	Employment probability	Vocational training: - men - women	(-) (0)	Simultaneous probit
Prey (1999)	Employment probability / wages	Vocational training	(-) / (0)	Simultaneous probit
Staat (1997)	Search duration / Employment probability / wages	Public sector training	(+) / (0) / (+)	Ordered probit, IV-estimation
Germany (East)				
Fitzenberger and Prey (1997)	Employment probability	Training outside the firm Training inside the firm	(+) (-)	Simultaneous probit
Hübner (1998)	Job security / Earnings	Off-the-job training	(+)	Comparison of: matching / OLS /Logit-ML
Fitzenberger and Prey (2000)	Employment / Earnings	Off-the-job training Public financed training	(+) (0) - (+)	Simultaneous probit
Pannenberg (1995, 1996)	Re-employment probability	Off-the-job training Off-the-job training	(+) (+)	Duration analysis
Lechner (1999b)	Income / Unemployment probability	Vocational training	(+) / (0)	Matching

Hübler (1994)	Search activity / Hours of work	Off-the-job training	(+)/(-)	Simultaneous probit
Kraus et al. (1998)	(Stable) Employment	Off-the-job training	(+)	Multinomial Logit
		Off-the-job training	(+)	
Bergemann et al. (2000)	Employment	Further training	(0)	Matching
Hujer and Wellner (2000a)	Unemployment duration	Public-sector sponsored vocational training	(0)	Duration analysis
Lechner (1999a)	Employment probability Earnings	Off-the-job training	(0)	Matching
			(0)	
Fitzenberger and Prey (1998)	Employment and wages	Training in the firm	(-)-(0)	Random-effects estimation, Matching
		Training outside the firm	(-)-(0)	
Lechner (2000)	Employment probabilities	Public-sector sponsored vocational training	(0)	Matching
* Effect on the outcome variable: (+) positive significant, (-) negative significant, 0 no significant effect.				

4. Policy implications: some guidance for evaluation and implementation practice

The last two chapters have presented several different evaluation methods and previous empirical findings for the evaluation of LMT in Germany and the EU. The three evaluation steps discussed (micro- and macroeconomic evaluation as well as CBA) should be seen as additional ingredients to a complete evaluation. Clearly, the first step of every evaluation, and therefore also the most dominant in existing literature, is based at individual level. This is easy to understand if we bring to mind that the most relevant question to be answered after the introduction of the new programme is whether the programme has the desired effects for the participating individual. Other effects, indirect ones on the non-participants or macroeconomic effects, have to be taken into account too, but they are usually assessed after the microeconomic evaluation. This chapter summarises and reviews the findings and aims to give guidance to policy-makers on how to evaluate and implement labour-market programmes. Therefore, focus is on the microeconomic approach, the choice of the estimation method, the problem of heterogeneity, data requirements, the importance of additional macroeconomic and CBA, the design of training programmes and the transferability of the findings to other social programmes.

4.1. Selection problem and the choice of estimation method

The fundamental evaluation problem and the risk of a selection bias are the first things to worry about in the microeconomic evaluation context. We would like to compare the outcome of the participating individual after the programme took place with the hypothetical outcome if they have not participated. Since we cannot observe the same individual simultaneously in both states (participation and non-partic-

ipation), the fundamental evaluation problem arises and, in order to estimate the true treatment effect, some identifying assumptions are required. We have presented several microeconomic evaluation estimators and the assumptions they impose in Section 2.2. We have also discussed how likely these assumptions are met in reality and showed how these estimators are implemented with a basic numerical example (Section 2.2.7). Of all the approaches presented, the matching estimator seems to be most favourable. The basic idea is to compare individuals who have the same characteristics. Ideally we have statistical twins and the only difference between them is participation in the programme. Therefore, we can interpret the difference in their outcomes as the average treatment effect of the programme. The implementation of the matching estimator relies on the assumption that all observable characteristics that influence the participation decision and the outcome variable can be controlled. Conditional to these characteristics, participation is independent of the outcomes and no selection bias should occur (conditional independence assumption). Thus it can be concluded that the matching estimator fully accounts for selection on observable characteristics. However, to implement it an informative dataset is needed which contains the variables from which they might influence the participation decision and the outcome variable ⁽²⁹⁾. If one suspects that unobservable characteristics (like the motivation of an individual) or unobserved variables (some relevant variables, like education or labour-market history are not in the dataset) drive the selection bias, an additional DID procedure might be useful. Even though it is hard to give a general recommendation in this context, it can be concluded that the more informative the dataset the more likely it is that the conditions for the matching estimator are met.

⁽²⁹⁾ In Section 4.3 we will discuss which variables that might be.

4.2. Heterogeneity

Another important issue for an evaluation is the problem of heterogeneity. Disregarding heterogeneity might lead to severe problems and biases for the estimation and interpretation of the results. Basically heterogeneity might arise on three levels (individual, programme, regional) which will be discussed later. First, an in-depth estimation of the effects of a training programme on the participating individual should regard the heterogeneity of the participants. For example, it makes a difference if one participant is long-term unemployed and another is unemployed only for a short time. This is because the effects for heterogeneous participants may differ significantly. To proceed with the example we assume that the effect of the programme is positive for the long-term unemployed but negative for the short-term unemployed. Taking the average over both groups would, therefore, lead to a biased estimation of the treatment effect. The positive effect for the long-term unemployed is underestimated and the negative effect for the short-term unemployed is overestimated. Clearly, this is a significant bias which has to be avoided by estimating the effects separately for the heterogeneous groups.

The second type of heterogeneity concerns the programmes. Programmes might differ for example in their duration or content. It is understandable that different programmes might have different effects. Comparing a two-week computer course with a six-month language course may be very misleading. Estimating the effect by taking the average over both programmes leads again to a biased estimate of the treatment effect. Finally, there is also a heterogeneity regarding the geographical distribution of the participants. The importance of local labour markets has been stressed in recent years. Different conditions in these local labour markets might affect the effectiveness of the programmes and the outcomes of participants and non-participants. In the context of the matching estimator, this means that comparing a participant from a booming region with a non-participant from a depressed region is not appropriate. But this might also extend to the programmes since a programme might work in a certain region but not in another one.

Bearing these considerations in mind, it should be emphasised that accounting for all three types of heterogeneity is crucial to obtaining a reliable and differentiated picture of the effects of LMT programmes. To do this certain data requirements have to be fulfilled and this is the topic of our next section.

4.3. Data requirements

The previous discussion has clarified the crucial role that datasets play in an evaluation study. To implement the matching estimator and to take account of the heterogeneity problem an informative dataset is needed. First individual information for participants and non-participants is required. To estimate the effects in heterogeneous subgroups of the treatment population (e.g. old vs. young participants) a sufficient number of treated individuals is needed. To find comparable control group members, the number of non-participants has to exceed the number of participants. Unfortunately there is no rule of thumb on how many control group members are needed to do this. Clearly the more comparable both groups are, the more likely it is that good matches will exist. For example, if the programme is addressed to the long-term unemployed and the control group is drawn out of long-term unemployed, the number of control units does not have to be much higher. On the other hand, if the control group is drawn randomly from the whole population (including short-term unemployed, employed, etc.) good matches are less likely and therefore the number of controls has to be higher. The type of explanatory variables which are needed depend on the programmes evaluated and should be based on a theoretical model (which variables influence the participation decision and the outcome variable?). The suggestions in this context are manifold. Sociodemographic variables like age, gender, education, children, etc., are essential as well as variables of the labour-market status (current job, industrial sector, function, high-skilled/low-skilled occupation). The importance of labour-market history as an explanatory variable for labour-market success today has been stressed in recent years. Therefore 'historical' information is used in most of the studies and is especially important if one suspects that there has been an Ashenfelter's dip. Finally, a good and reliable

outcome measure needs to be available too. The outcome measure should correspond to the aims of the programme (e.g. if the programme is intended to increase the employment prospects of individuals we need the employment status after the programme). It should be traced a sufficient period of time after the programme ends to draw conclusions for the short-, medium- and long-term effects of the programme. The reliability of the outcome measure is very important since the estimation of the effects depends directly on it.

One well-known problem for evaluators is that they are not included in the creation of the datasets used later on for the programme evaluation. In this situation one can either rely on already existing datasets or build up new ones. Using existing datasets, for example surveys like the German Socioeconomic Panel or administrative records collected by governmental agencies, has the major advantage of low costs. There are, however, also several drawbacks which make the interpretation of the results problematic (e.g. Heckman et al., 1999). First, such datasets are not focused on participants in certain programmes, i.e. the number of observations in the treatment group will probably be quite small. Additionally, even if information on treated individuals is available, it is very often difficult to construct an appropriate control group mirroring the treatment group in all characteristics. It may also be difficult to implement the matching estimator if not enough explanatory variables are available. Another possible factor of distortion is given when the outcome variable of interest is not covered in the dataset and has to be approximated. The major disadvantage of building up new datasets (either by collecting a new survey or merging the information from administrative records, etc.) is the high costs. Besides that, the advantages are numerous. The researcher has complete control over the information collected (important variables, outcome variable, etc.) in the group of treated and control individuals. The decisive objective of the programme can be taken into account (e.g. not only the employment status but also the hours worked and/or the wage) and the conditions for the implementation of the appropriate econometric method are more likely to hold. The ideal evaluation dataset is a panel dataset for participants and non-participants that starts some time before the

programme begins and that traces the individuals for a certain time period. The necessary length of this time period depends on the views of the policy-makers on how long the effects last.

4.4. Macroeconometric analysis and cost-benefit analysis

Evaluating the impacts of labour-market programmes at an individual level is an important topic but can only serve as a first step in a comprehensive evaluation. This microeconomic approach has to be complemented by two additional steps. The first consists of an assessment of the indirect effects which can emerge at a macroeconomic level. Examples of such indirect spillover effects include dead-weight losses, substitution or displacement effects. Since these effects can counteract the actual goals and objectives of a programme, it is crucial to take them into account. The appropriate method in this context is dynamic panel models as they allow for persistent patterns in the labour market. The data requirements for their implementation are usually substantially lower compared to the microeconomic analysis. Aggregate (either on national or regional) data is required and variables of interest include: (un)employed, programme participants, spending on ALMP and PLMP, vacancies, etc. Particularly interesting in this context are aggregated flow data (from programmes into (un)employment and vice versa), for example to estimate the relationship between the number of unemployed which are put back into work through programmes. An additional methodological problem which has to be taken into account in this context is a potential simultaneity between the spending on labour-market programmes and unemployment. Clearly, in macroeconomic analysis it is harder to take account of the afore-mentioned heterogeneity, since individual characteristics are usually not available.

Once all micro- and macro-economic effects have been evaluated, confrontation between these estimated effects and associated costs is imperative, i.e. the next logical step is to conduct a CBA. We sketched the proceeding for a CBA in one of the previous sections. One of the major problems was a complete gathering of all benefits of labour-market programmes. Surely the effects

which are estimated typically by microeconomic methods, for example improvement in labour-market prospects, are not exhaustive. Instead various additional indirect effects have to be taken into account. Typical examples in this context include the reduction in criminal activities or equity aims. Once this enumeration is finished the next, and even more difficult, step is measuring these effects in monetary terms. If there is no clear guidance from doing so, at least a cost-effectiveness analysis should be conducted. We have given a numerical example for such a cost-effectiveness analysis in Section 2.4.7.

4.5. The design of training programmes: some suggestions

Since the importance of heterogeneity has been stressed in the previous sections, can it be concluded from empirical findings which programme designs are most effective? Naturally this question can only be answered with evaluation studies that compare different types of vocational training programmes. Since macroeconomic evaluations do not distinguish between different types of vocational training programmes, the following statements will rely solely on the review of microeconomic studies. It should be noted, however, that the suggestions we make may not be applicable across all occurrences. The success of every programme also depends on the (labour-market) situation in the implementing regions and therefore a rule of thumb is not available.

The major objective of LMT programmes is to reintegrate the unemployed into regular employment. In order to accomplish this objective the choice between on-the-job and off-the-job training is crucial. A straightforward argument would be that on-the-job training is more favourable, since such a programme does not only provide training but also imparts work experience. Another point might be that the immediate application of the learned knowledge should make it easier for participants to transfer the learned knowledge into their work. Considering the empirical results in respect of this context, no clear picture emerges. The study from Hujer et al. (1999c) suggests a more favourable effect for

off-the-job training programmes; that on-the-job training programmes are too specific and only of use in the same firm. In contrast to this finding the majority of studies find that on-the-job training is more effective. The study from Johansson and Martinson (2000) suggests that a narrow contact between participants and firms is beneficial. The subject of the programme should also be considered in this context. General courses, like language or basic computer courses, may work better as off-the job training programmes, but if a training programme is aimed at a specific skill that is narrowly associated with practical work, on-the-job training seems to be more appropriate.

Another question might be how long the duration of a training programme should be in order to be most efficient. As a first guess it could be argued that a longer programme means more comprehensive training. However, participation in a training programme is mostly associated with reduced search activity and therefore the longer programme duration may thwart the effects. The locking-in effect in particular has become a major issue in empirical studies. The problem here is that programmes associated with a full-time engagement do not allow any time for active job search, the benefit of new knowledge is opposed to reduced search activity. In order to avoid this negative effect, vocational training programmes should be designed in a way that enough time remains for active search. Additionally, active search should be actively promoted. Compensation for the participants is relevant in this context too. If the compensation is too high, the incentive to search for a regular job is weakened. Therefore, compensation should lie significantly under the earnings associated with a regular employment.

A further concern associated with training programmes seems to be that, although these programmes are designed for problem groups, an often found phenomenon is that the organising institution selects the participants in a way that the success of the programme is artificially improved. In order to avoid this 'creaming' of participants, it could be useful to define the eligibility criteria for a programme participation in a fairly explicit way; the length of the unemployment spell, as well as the basic skill level, should serve as a major criterion. As the Winter-Ebmer study (2001) has shown, vocational training

programmes that are targeted to an explicit group can be quite successful. An additional advantage of explicit targeting is that the content of the training can be adjusted to the needs of the participants in a more efficient way.

A final issue is that, in some European countries, participation in a training programme is essential to preserve entitlement to unemployment benefits. Although this helps to avoid a misuse of the unemployment benefit system, it may be contradictory for the success of the training programme, since training cannot be effective if the participants are forced to participate. If there is the objective to inhibit misuse of the benefit system, job creation schemes seem to be more convenient, since they are of value also for the whole society.

The last point shows that the effectiveness of vocational training programmes does not only depend on the design and implementation of the programmes itself, but also on the general framework of labour-market policy. There is a strong interdependence between the unemployment benefit system, i.e. the passive labour-market policy, and vocational training, i.e. ALMP. Therefore, general recommendations are hard to make and the design has to be adjusted to the local situation.

4.6. Transferability to other social programmes

An important question which remains unanswered regards the transferability of the methods presented to other social programmes. Even though, in principle, transfer is possible, certain restrictions and possible problems will be considered in this section. Although able to reveal the impacts at individual and macroeconomic levels, micro- and macroeconomic evaluation methods do not cover the whole scope of effects associated with social programmes. The advantage of econometric methods lies in assessing the quantitative effects of social programmes. In evaluating LMT, for example, they are able to uncover whether the employment chances of the participants or the (un)employment situation in the economy are affected. Social programmes, however, might also involve various other more qualitative aims and objectives. These goals (e.g. well-being, health, criminal records) are often hard to measure and therefore hard to evaluate in

the presented framework. The time horizon of the programme might also be a problem. The matching estimator identifies the true treatment effect by controlling for other characteristics which influence the outcome variable. Naturally, the longer the time horizon of the programme or the evaluation itself, the harder it is to control for all influences. After 10 years, for example, it is reasonable to assume that other factors which have nothing to do with the treatment are influencing the outcome variable. Therefore, the applicability of the estimator is questionable in this situation. Furthermore, all presented estimators have some data requirements (Section 4.3) which have to be met for their implementation. If these requirements are not fulfilled, their application fails. Another point concerns the choice of the control group. In the methodological discussion we introduced the control group as a group of non-participants. This is problematic if no non-participants are available, for example if all the unemployed have to participate in a measure after a certain period of unemployment or if comparison of the effects of different vocational training programmes is required. This problem is not so problematic and the framework can still be applied. The important difference is the different interpretation of the results. For example, if we are interested in comparing two training programmes we can use the participants of the second training group as controls. So, instead of comparing participation in a programme with non-participation we compare the effect of programme A relative to the effect of programme B. If we only have one identical programme in which all individuals have to participate, it might be interesting to compare individuals who participate after three months of unemployment with those who participate after two years of unemployment. This could lead to conclusions on when it is best to direct the unemployed into programmes. To emphasise this point: the control group does not necessarily have to be a group of non-participants but might also be a group of participants in other programmes or a group within the same programme but with other entry characteristics. Attention has to be given, in this case, to interpretation of the effects, since they are no longer measured relative to non-participation.

5. Summary and conclusions

The recent growth in public spending on social programmes, together with tight government budgets, has increased the demand for evaluation of these programmes. In consequence, several evaluation methods and procedures have been developed. With this contribution we have tried to give an overview of the necessary steps in such analysis. The ideal evaluation process can be viewed as a series of three steps. First, the impacts of the programme on the individual should be estimated. Second, it should be examined whether or not the estimated impacts are large enough to yield net social gains. Finally, it should be determined if this is the best outcome that could have been achieved for the money spent (Fay, 1996). For these reasons, we concentrated on evaluation of economic goals with econometric techniques.

The first step of an evaluation is concerned with the individual level, i.e. the researcher has to analyse if the programme under consideration has any effect at all on the individual. This means answering the question of how the participating individual would have behaved under the hypothetical situation of non-participation. Has the individuals who participated in this programme improved in contrast to a situation without participation? The fundamental evaluation problem which arises in this context is due to the fact that we can never observe an individual in both states, i.e. we have a counterfactual situation. We have presented several econometric estimators which simulate this hypothetical outcome. These estimators rest on a number of generally untestable assumptions which were discussed critically to determine their major advantages and disadvantages. We also assessed how likely it is that these assumptions are met in practice. Furthermore, we implemented a numerical example to show how the estimators are implemented technically and also discussed their data requirements.

This microeconomic evaluation, is only a first step towards an overall assessment of the social programme. In this contribution we have

also sketched how the next steps should look in an evaluation study. We pointed out the possible existence of indirect and secondary effects which might counteract the actual aims and objectives of the programme. In this context the most important effects are dead-weight losses, substitution and displacement effects. These effects cannot be evaluated at an individual level, making an additional assessment on an aggregate level necessary. We have proposed macroeconomic methods, like augmented matching functions which can be used to take such indirect effects into account. Instead of looking at the effect on individual performance, we would like to know if the ALMP represent a net gain to the whole economy. Important methodological issues in a macroeconomic evaluation are the specification of the empirical model, which should always be based on an appropriate theoretical framework and the simultaneity problem of ALMP which has to be solved. Having identified all potential quantitative effects of a programme, whether they are direct or indirect, the final step consists of augmenting these effects with additional, more qualitative impacts of the programme. These are harder to measure but nevertheless important for an overall assessment. The defined benefits of the programme can be contrasted in a CBA with the costs caused by the programme in order to assess the overall net benefit.

Having discussed the methodological issues related to the three evaluation steps, we presented empirical findings from micro- and macroeconomic evaluations in Europe, with a particular focus on Germany. Training programmes seem to have positive effects on the individuals in most of the studies and perform better than alternative labour-market programmes, such as job creation schemes. Finally, we used the results from our methodological discussion and the empirical findings to draw some implications for policy and evaluation practice. We focused on the choice of the appropriate estimation method, data requirements, the problem of heterogeneity in evaluation

analysis, the successful design of training programmes and the transferability of our findings (for ALMP) to other social programmes. It was our aim to provide the interested reader with the methodological foundations of econometric evalu-

ation analysis and present some easily understandable numerical examples. In doing so, we wanted to give some guidance for the evaluation and implementation practice of ALMP and, to a certain extent, also other social programmes.

List of abbreviations

ALMP	Active labour-market policies
BAE	Before-after estimator
CBA	Cost-benefit analysis
CEA	Cost-effectiveness analysis
CIA	Conditional independence assumption
CSE	Cross-section estimator
DID	Difference-in-differences
EGLMM	East German Labour Market Monitor
EU	European Union
GDP	Gross domestic product
GNP	Gross national product
GSOEP	German socioeconomic panel
JCS	Job creation scheme
KM	Kernel matching
LMT	Labour-market training
LTU	Long-term unemployment
NN	Nearest neighbour
NPV	Net present value
PLMP	Passive labour-market policies
SAM	Social assistance measure
SUTVA	Stable unit treatment value assumption
SwIT	Sweden Information Technology

Annex: Tables

Table A1: **Standardised unemployment rates in the EU (as a percentage of total labour force)**

(%)

[illegible]

(%)

	1985	1986	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	Avg.	Rel. Change 1985-2000
Austria	0.08	0.11	0.10	0.11	0.08	0.10	0.11	0.12	0.15	0.17	0.15	0.19	0.17	0.13	113%
Belgium	0.19	0.21	0.21	0.22	0.23	0.27	0.28	0.28	0.28	0.26	0.25	0.24	0.24	0.25	26%
Denmark	. .	0.43	0.27	0.35	0.39	0.47	0.69	0.98	1.07	0.93	0.97	0.99	0.85	0.72	98%
Finland	0.26	0.27	0.25	0.33	0.46	0.47	0.46	0.44	0.55	0.53	0.44	0.40	0.30	0.42	15%
France	0.26	0.27	0.33	0.35	0.38	0.43	0.41	0.38	0.36	0.34	0.31	0.28	0.25	0.35	-4%
Germany	0.2	0.24	0.38	0.45	0.63	0.55	0.41	0.38	0.45	0.35	0.34	0.35	0.34	0.42	70%
Greece	0.02	0.03	0.16	0.21	0.12	0.08	0.06	0.13	0.09	0.06	0.21	. .	. .	0.12	950%
Ireland	0.63	0.57	0.47	0.24	. .	. .	0.23	0.22	0.21	. .	. .	. .	. .	0.27	-66%
Italy	. .	. .	. .	0.02	0.02	0.01	0.01	0.01	0.01	0.12	0.15	0.12	. .	0.05	583%
Netherlands	0.13	0.19	0.27	0.29	0.39	0.37	0.42	0.39	0.34	0.35	0.31	0.34	0.30	0.34	131%
Portugal	. .	0.18	0.12	0.17	0.26	0.25	0.20	0.23	0.29	0.27	0.29	0.29	0.15	0.23	-17%
Spain	0.02	0.07	0.16	0.17	0.10	0.11	0.22	0.31	0.33	0.15	0.17	0.12	0.15	0.18	650%
Sweden	0.5	0.5	0.53	1.01	1.09	0.76	0.77	0.55	. .	0.42	0.45	0.48	0.31	0.64	-38%
UK	0.07	0.07	0.21	0.15	0.16	0.15	0.14	0.10	0.09	0.07	0.07	0.05	0.05	0.11	-29%
Average	0.21	0.24	0.26	0.29	0.33	0.31	0.32	0.32	0.32	0.31	0.32	0.32	0.28	0.30	32%

. . Missing values
Source: Employment Outlook (OECD, 2002); Martin, 1998

References

- Aakvik, A.; Heckman, J. J.; Vytlačil, E. J. *Treatment effects for discrete outcomes when responses to treatment vary among observationally identical persons: an application to Norwegian vocational rehabilitation programmes*. Cambridge: National Bureau of Economic Research – NBER, 2000 (Technical Working Paper, 262).
- Ashenfelter, O. Estimating the effects of training programmes on earnings. In: *Review of Economics and Statistics*, 1978, Vol. 60, p. 47-57.
- Ashenfelter, O.; Card, D. Using the longitudinal structure of earnings to estimate the effects of training programmes. In: *Review of Economics and Statistics*, 1985, Vol. 66, p. 648-660.
- Auer, P.; Kruppe, T. *Monitoring of labour market policy in the EU member states*. Berlin: Reihe, 1996 (Serie: Wissenschaftszentrum Berlin für Sozialforschung. Discussion paper FS 1 No 96-202).
- Augurzky, B. *Optimal full matching: an application using the NLSY79*. Heidelberg: University of Heidelberg, Department of Economics, 2000. (Discussion Paper, No 310).
- Bergemann, A. et al. *Multiple active labor market policy participation in East Germany: an assessment of outcomes*. Halle: Institute of Economic Research, 2000 (Working Paper).
- Bijwaard, G.; Ridder, G. *Correcting for selective compliance in a reemployment bonus experiment*. Baltimore: John Hopkins University, 2000 (Working Paper).
- Boardman, A. E. et al. *Cost-benefit analysis: concepts and practice*. New Jersey: Prentice Hall, 2001.
- Bonnal, L.; Fougère, D.; Sérandon, A. Evaluating the impact of French employment policies on individual labour market histories. In: *Review of Economic Studies*, 1997, p. 683-713.
- Brent, R. J. *Applied cost-benefit analysis*. Vermont: Edward Elgar Publishing Limited, 1996.
- Brodsky, T.; Crepon, B.; Fougère, D. Using matching estimators to evaluate alternative youth employment programmes: evidence from France, 1986-1988. In: Lechner, M.; Pfeiffer, F. (eds) *Econometric evaluation of labour market policies*. Heidelberg: Physika Verlag, 2001 (ZEW Economic Studies, Vol. 13).
- Burtless, G. The case for randomized field trials in economic and policy research. In: *Journal of Economic Perspectives*, 1995, Vol. 9, p. 63-84.
- Burtless, G.; Orr, L. Are classical experiments needed for manpower policy? In: *Journal of Human Resources*, 1986, Vol. 21, p. 606-640.
- Büttner, T.; Prey, H. Does active labour-market policy affect structural unemployment? An empirical investigation for West Germany regions, 1986-1993. In: *Zeitschrift für Wirtschafts- und Sozialwissenschaften*, 1998, Vol. 188, p. 389-413.
- Calmfors, L. Active labour market policy and unemployment: a framework for the analysis of crucial design features. In: *OECD Economic Studies*, 1994, Vol. 22, p. 7-47.
- Calmfors, L.; Forslund, A.; Hemström, M. *Does active labour market policy work? Lessons from the Swedish experiences*. Munich: CESifo, 2002 (Working Paper series, No 675).
- Calmfors, L.; Skedinger, P. Does active labour market policy increase employment? Theoretical considerations and some empirical evidence from Sweden. In: *Oxford Review of Economic Policy*, 1995, No 11.
- Card, D. *Reforming the financial incentives of the welfare system*. Bonn: IZA, 2000 (Discussion Paper, No 172).
- Cochrane, W.; Rubin, D. Controlling bias in observational studies. In: *Sankhya*, 1973, Vol. 35, p. 417-446.
- Cockx, B.; Bardoulat, I. *Vocational training: does it speed up the transition rate out of unemployment?* Amsterdam: Tinbergen Institute, 2000 (Discussion Paper No 9932).
- Cockx, B.; Van Der Linden, B.; Karaa, A. Active labour market policy and job tenure. In: *Oxford Economic Papers*, 1998, Vol. 50, p. 685-708.
- Delander, L.; Niklasson, H. Cost-benefit analysis. In: Schmid, G.; O'Reilly, J.; Schönemann, K. (eds) *Handbook of international labour market policy and economics*. Cheltenham: Edward Elgar, 1996.

- DiNardo, J.; Fortin, N.; Lemieux, T. Labor market institutions and the distribution of wages, 1973-1992: a semiparametric approach. In: *Econometrica*, 1996, Vol. 64, p. 1001-1045.
- Dolton, P.; O'Neill, D. Unemployment duration and the restart effect: some experimental evidence. In: *Economic Journal*, 1996, Vol. 106, p. 387-400.
- Fay, R. *Enhancing the effectiveness of active labour market policies: evidence from programme evaluations in OECD countries*. Paris: OECD, 1996 (Labour Market and Social Policy Occasional Papers).
- Firth, D.; Payne, C.; Payne, J. Efficacy of programmes for the unemployed: discrete time modelling of duration data from a matched comparison study. In: *Journal of the Royal Statistical Society*, 1999, Vol. 162, p. 111-120.
- Fitzenberger, B.; Prey, H. Assessing the impact of training on employment. In: *ifo-Studien*, 1997, Vol. 43, p. 71-116.
- Fitzenberger, B.; Prey, H. Evaluating public sector sponsored training in East Germany. In: *Oxford Economic Papers*, 2000, p. 497-520.
- Fitzenberger, B.; Prey, H. Beschäftigungs- und Verdienstwirkungen von Weiterbildungsmaßnahmen im ostdeutschen Transformationsprozeß: Eine Methodenkritik. In: Pfeiffer, F.; Pohlmeier, W. (eds) *Qualifikation, Weiterbildung und Arbeitsmarkterfolg*. Baden-Baden: Nomos-Verlag, 1998 (ZEW-Wirtschaftsanalysen, Vol 31).
- Gerfin, M.; Lechner, M. *Microeconomic evaluation of the active labour market policy in Switzerland*. Mannheim: ZEW, 2000 (Discussion Paper, No 00-24).
- Gritz, M. The impact of training on the frequency and duration of employment. In: *Journal of Econometrics*, 1993, Vol. 57, p. 21-51.
- Hagen, T.; Steiner, V. *Von der Finanzierung der Arbeitslosigkeit zur Förderung von Arbeit – Analysen und Empfehlungen zur Arbeitsmarktpolitik in Deutschland*. Baden-Baden: Nomos Verlagsgesellschaft, 2000.
- Heckman, J.; Hotz, J. Choosing among alternative nonexperimental methods for estimating the impact of social programmes: the case of manpower training. In: *Journal of the American Statistical Association*, 1989, Vol. 84, p. 862-880.
- Heckman, J. et al. Sources of selection bias in evaluating social programmes: an interpretation of conventional measures and evidence on the effectiveness of matching as a program evaluation method. In: *Proceedings of the National Academy of Sciences*, 1996, Vol. 93, p. 13416-13420.
- Heckman, J. et al. Characterizing selection bias using experimental data. In: *Econometrica*, 1998a, Vol. 66, p. 1017-1098.
- Heckman, J.; Ichimura, H.; Todd, P. Matching as an econometric evaluation estimator: evidence from evaluating a job training programme. In: *Review of Economic Studies*, 1997, Vol. 64, p. 605-654.
- Heckman, J.; Ichimura, H.; Todd, P. Matching as an econometric evaluation estimator. In: *Review of Economic Studies*, 1998b, Vol. 65, p. 261-294.
- Heckman, J.; LaLonde, R.; Smith, J. The economics and econometrics of active labor market programmes. In: Ashenfelter, O.; Card, D. (eds) In: *Handbook of Labor Economics Vol. III*. Amsterdam: Elsevier, 1999, p. 1865-2097.
- Heckman, J.; Robb, R. Alternative models for evaluating the impact of interventions. In: Heckman, J.; Singer, B. (eds) *Longitudinal analysis of labor market data*, Cambridge: Cambridge University Press, 1985, p. 156-245.
- Heckman, J.; Singer, B. A method for minimizing the impact of distributional assumptions in econometric models for duration data. In: *Econometrica*, 1984, Vol. 52, p. 271-320.
- Heckman, J.; Smith, J. Assessing the case for social experiments. In: *Journal of Economic Perspectives*, 1995, Vol. 9, p. 85-110.
- Holland, P. Statistics and causal inference. In: *Journal of American Statistical Association*, 1986, Vol. 81, p. 945-970.
- Holmlund, B.; Linden, J. Job matching, temporary public employment and equilibrium unemployment. In: *Journal of Public Economics*, 1993, Vol. 51, p. 329-343.
- Hübler, O. Berufliche Weiterbildung und Umschulung in Ostdeutschland – Erfahrungen und Perspektiven. In: Pfeiffer, F.; Pohlmeier, W. (eds) *Qualifikation, Weiterbildung und Arbeitsmarkterfolg*. Baden-Baden: Nomos-Verlag, 1998 (ZEW-Wirtschaftsanalysen, Vol. 31).
- Hübler, O. Weiterbildung, Arbeitsplatzsuche und individueller Beschäftigungsumfang – Eine ökonometrische Untersuchung für Ostdeutschland. In: *Zeitschrift für Wirtschafts- und Sozialwissenschaften*, 1994, Vol. 114, p. 419-447.

- Hujer, R.; Caliendo, M. Evaluation of active labour market policy – methodological concepts and empirical estimates. In: Becker, I. et al. (eds) *Soziale Sicherung in einer dynamischen Gesellschaft*. Campus Verlag, 2001, p. 583-617.
- Hujer, R.; Maurer, K. O.; Wellner, M. Kurz- und langfristige Effekte von Weiterbildungsmaßnahmen auf die Arbeitslosigkeitsdauer in Westdeutschland. In: Pfeiffer, F.; Pohlmeier, W. (eds) *Qualifikation, Weiterbildung und Arbeitsmarkterfolg*. Baden Baden: Nomos-Verlag, 1998, p. 197-221 (ZEW-Wirtschaftsanalysen Band 31).
- Hujer, R.; Maurer, K. O.; Wellner, M. Estimating the effect of vocational training on unemployment duration in West Germany: a discrete hazard rate model with instrumental variables. In: *Jahrbücher für Nationalökonomie und Statistik*, 1999a, Vol. 218/5+6, p. 619-646.
- Hujer, R.; Maurer, K. O.; Wellner, M. The effects of public sector sponsored training on unemployment duration in West Germany: a discrete hazard rate model based on a matched sample. In: *ifo-Studien*, 1999b, Vol. 3, p. 371-410.
- Hujer, R.; Maurer, K. O.; Wellner, M. Analyzing the effects of on-the-job vs. off-the-job training on unemployment duration in West Germany. In: Bellmann, L.; Steiner, V. (eds) *Panelanalysen zu Lohnstruktur, Qualifikation und Beschäftigungsdynamik*. IAB – Beiträge zur Arbeitsmarkt- und Berufsforschung 229, 1999c, p. 203-237.
- Hujer, R.; Wellner, M. *The effects of public sector sponsored training on individual employment performance in East Germany*. Bonn: IZA, 2000a (Discussion Paper, No 141).
- Hujer, R.; Wellner, M. *Berufliche Weiterbildung und individuelle Arbeitslosigkeitsdauer in West- und Ostdeutschland: Eine mikroökonomische Analyse*, 2000b (Discussion Paper).
- Jackman, R.; Pissarides, C.; Savouri, S. Labour market policies and unemployment in the OECD. In: *Economic Policy*, 1990, Vol. 5, p. 450-490.
- Jensen, P.; Nielsen, M. S.; Rosholm, M. *The effects of benefit, incentives, and sanctions on youth unemployment*. Aarhus: Center of Labour Market and Social Research – CLS, 1999 (Working Paper, No 99-05).
- Johansson, P.; Martinson, S. *The effect of increased employer contacts within a labour market training program*. Uppsala: IFAU – Office of Labour Market Policy Evaluation, 2000 (Working Paper, No 10).
- Kluve, J.; Lehmann, H.; Schmidt, C. Active labour market policies in Poland: human capital enhancement, stigmatization, or benefit churning? In: *Journal of Comparative Economics*, 1999, Vol. 27, p. 61-89.
- Koenker, R.; Biliass, Y. Quantile regression for duration data: a reappraisal of the Pennsylvania reemployment bonus experiments. In: Fitzenberger, B.; Koenker, R. Machado, J. A. F. (eds) *Studies in empirical economics: economic applications of quantile Regression*, 2002, p. 199-220.
- Kraus, F.; Puhani, P. A.; Steiner, V. *Do public works programmes work? Some unpleasant results from the East German experience*. Mannheim: ZEW, 1998 (Discussion Paper, No 98-07).
- Kromrey, H. Evaluation – ein vielschichtiges Konzept. In: *Sozialwissenschaften und Berufspraxis*, 2001, Vol. 2, p. 105-132.
- Lalive, R.; Van Ours, J. C.; Zweimüller, J. *The impact of active labor market programmes and benefit entitlement rules on the duration of unemployment*. Bonn: IZA, 2000. (Discussion Paper No 149).
- LaLonde, R. Evaluating the econometric evaluations of training programmes with experimental data. In: *The American Economic Review*, 1986, Vol. 76, p. 604-620.
- LaLonde, R. The promise of public-sector sponsored training programmes. In: *Journal of Economic Perspectives*, 1995, Vol. 9, p. 149-168.
- Larsson, L. *Evaluation of Swedish youth labour market programmes*. Uppsala: IFAU, 2000 (Working Paper, 2000:1).
- Layard, R.; Nickell, S.; Jackman, R. *Unemployment: macroeconomic performance and the labour market*. Oxford: Oxford University Press, 1991.
- Lechner, M. Mikroökonomische Evaluationsstudien: Anmerkungen zu Theorie und Praxis. In: Pfeiffer, F.; Pohlmeier, W. (eds) *Qualifikation, Weiterbildung und Arbeitsmarkterfolg*. Baden-Baden: Nomos-Verlag, 1998. (ZEW-Wirtschaftsanalysen, Vol. 31).
- Lechner, M. Earnings and employment effects of continuous off-the-job training in East Germany after unification. In: *Journal of Business Economic Statistics*, 1999c, Vol. 17, p. 74-90.
- Lechner, M. An evaluation of public sector sponsored continuous vocational training programmes in East Germany. In: *Journal of Human Resources*, 2000a, p. 347-375.

- Lechner, M. The effects of enterprise-related continuous vocational training in East Germany on individual employment and earnings. In: *Annals d'Economie et de Statistique*, 1999d, Vol. 55-56, p. 97-128.
- Long, D. A.; Mallar, C. D.; Thornton, C. V. B. Evaluating the benefits and costs of the job corps. In: *Journal of Policy Analysis*, 1981, Vol. 1(1), p. 55-76.
- Lubyova, M.; Van Ours, J. Effects of active labour market programmes on the transition rate from unemployment into regular jobs in the Slovak Republic. In: *Journal of Comparative Economics*, 1999, Vol. 27, p. 90-112.
- Martin, J. *What works among active labour market policies: evidence from OECD countries*. Paris: OECD, 1998 (Occasional Papers).
- O'Connell, P.; McGinnity, F. What works, who works? The employment and earnings effects of active labour market programmes among young people in Ireland. In: *Work, Employment and Society*, 1997, Vol. 11, No 4, p. 639-661.
- OECD. *Economic Survey Italy 1986/87*. Paris: OECD, 1987.
- OECD. *Economic Survey Belgium 1987/88*. Paris: OECD, 1988a.
- OECD. *Economic Survey Denmark 1987/88*. Paris: OECD, 1988b.
- OECD. *Economic Survey Finland 1987/88*. Paris: OECD, 1988c.
- OECD. *Economic Survey France 1987/88*. Paris: OECD, 1988d.
- OECD. *Economic Survey Ireland 1987/88*. Paris: OECD, 1988e.
- OECD. *Economic Survey Portugal 1987/88*. Paris: OECD, 1988f.
- OECD. *Economic Survey Spain 1987/88*. Paris: OECD, 1988g.
- OECD. *Economic Survey Austria 1988/89*. Paris: OECD, 1989a.
- OECD. *Economic Survey United Kingdom 1988/89*. Paris: OECD, 1989b.
- OECD. *Economic Survey Greece 1989/90*. Paris: OECD, 1990a.
- OECD. *Economic Survey Germany 1989/90*. Paris: OECD, 1990b.
- OECD. *Economic Survey Netherlands 1989/90*. Paris: OECD, 1990c.
- OECD. *Economic Survey Sweden 1990/91*. Paris: OECD, 1990d.
- OECD. *Employment Outlook*. Paris: OECD, 1991.
- OECD. *Employment Outlook*. Paris: OECD, 1993.
- OECD. *Employment Outlook*. Paris: OECD, 2001.
- OECD. *Employment Outlook*. Paris: OECD, 2002.
- Pannenberg, M. *Weiterbildungsaktivitäten und Erwerbsbiographie – Eine empirische Analyse für Deutschland*. Frankfurt: Campus-Verlag, 1995.
- Pannenberg, M. *Zur Evaluation staatlicher Qualifizierungsmaßnahmen in Ostdeutschland: das Instrument Fortbildung und Umschulung (FuU)*. Halle: Institute for Economic Research, 1996 (Discussion Paper, No 38).
- Pannenberg, M.; Schwarze, J. Labour market slack and the wage curve. In: *Economics Letters*, 1998, Vol. 58, p. 351-354.
- Payne, J.; Lissenburg, S.; Withe, M. *Employment training and employment action: an evaluation by the matched comparison method*. London: Policy Studies Institute, 1996.
- Pohnke, C. *Wirkungs- und Kosten-Nutzen Analysen: Eine Untersuchung von Maßnahmen der aktiven Arbeitsmarktpolitik am Beispiel kommunaler Beschäftigungsprogramme*. Frankfurt: Lang, 2001.
- Prey, H. *Wirkungen staatlicher Qualifizierungsmaßnahmen. Eine empirische Untersuchung für die Bundesrepublik Deutschland*. Bern, Stuttgart, Wien: Paul Haupt-Verlag, 1999.
- Prey, H. *Wirkungen von Maßnahmen staatlicher Arbeitsmarkt- und Beschäftigungspolitik*. Konstanz: University of Konstanz, 1997 (CILE Discussion Paper, No 45).
- Richardson, K.; Van den Berg, G. *Swedish labour market training and the duration of unemployment*, 2001 (Working Paper, forthcoming).
- Rosenbaum, P.; Rubin, D. The central role of the propensity score in observational studies for causal effects. In: *Biometrika*, 1983, Vol. 70, p. 41-50.
- Rosenbaum, P.; Rubin, D. The bias due to incomplete matching. In: *Biometrics*, 1985a, Vol. 41, p. 103-116.
- Rosenbaum, P.; Rubin, D. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. In: *The American Statistician*, 1985b, Vol. 39, p. 33-38.
- Roy, A. Some thoughts on the distribution of earnings. In: *Oxford Economic Papers*, 1951, Vol. 3, p. 135-145.
- Rubin, D. Estimating causal effects to treatments in randomised and nonrandomised studies. In:

- Journal of Educational Psychology*, 1974, Vol. 66, p. 688-701.
- Rubin, D. Assignment to treatment group on the basis of a covariate. In: *Journal of Educational Studies*, 1977, Vol. 2, p. 1-26.
- Rubin, D. Using multivariate matched sampling and regression adjustment to control bias in observational studies. In: *Journal of the American Statistical Association*, 1979, Vol. 74, p. 318-328.
- Rubin, D. Comment on Basu, D. Randomization analysis of experimental data: The Fisher randomization test. In: *Journal of the American Statistical Association*, 1980, Vol. 75, p. 591-593.
- Schmidt, C. *Knowing what works: the case for rigorous program evaluation*. Bonn: IZA, 1999 (Discussion Paper, No 77).
- Schmid, G.; Speckesser, S.; Hilbert, C. Does active labour market policy matter? An aggregate impact analysis for Germany. In: *Labour market policy and unemployment. Evaluation of active measures in France, Germany, The Netherlands, Spain and Sweden*, Cheltenham: Edward Elgar, 2000.
- Scriven, M. *Evaluation thesaurus* (Fourth edition). Newbury Park, CA: Sage Publications, 1991.
- Sianesi, B. *An evaluation of the active labour market programmes in Sweden*. Uppsala: IFAU, 2001 (Working Paper, 2001:5).
- Smith, J. Evaluating active labour market policies: lessons from North America. In: *Mitteilungen aus der Arbeitsmarkt und Berufsforschung, Schwerpunkttheft: Erfolgskontrolle aktiver Arbeitsmarktpolitik*, 2000, Vol. 3, p. 345-356.
- Smith, J.; Todd, P. *Does matching overcome LaLonde's critique of nonexperimental estimators*. University of Western Ontario, University of Pennsylvania, 2000. (Working Paper).
- Staat, M. *Empirische Evaluation von Fortbildung und Umschulung*. Baden-Baden: Nomos-Verlag, 1997 (Schriftenreihe des ZEW, 21).
- Steiner, V.; Hagen, T. Was kann die Aktive Arbeitsmarktpolitik in Deutschland aus der Evaluationsforschung in anderen europäischen Ländern lernen? In: *Perspektiven der Wirtschaftspolitik 2000*, 2002, Vol. 3, No 2, p. 189-206.
- Steiner, V. et al. *Strukturanalyse der Arbeitsmarktentwicklung in den neuen Bundesländern*. Baden-Baden: Nomos Verlag, 1998.
- Van Ours, J. *Do active labour market policies help unemployed workers to find and keep regular jobs?* Tilburg: Center for Economic Research, 2000 (Discussion Paper, No 0010).
- Winter-Ebmer, R. *Evaluating an innovative redundancy-retraining project: the Austrian Steel Foundation*. London: CEPR – Centre for Economic policy research 2001 (Discussion Paper, DP2776).
- Worthen, B. R.; Sanders, J. R.; Fitzpatrick, J. L. *Program evaluation: alternative approaches and practical guidelines* (Second edition). New York: Longman, 1997.
- Zweimüller, J.; Winter-Ebmer, R. Manpower training programmes and employment stability. In: *Economica*, 1996, Vol. 63, p. 113-130.

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 ⁽¹⁾

Abstract

This study follows the international trend in evaluation approaches in vocational education and training (VET), moving away from single projects towards policy evaluations. It examines the contribution of research and a number of examples from transition countries in central and south-east European countries with a view to helping clarify underlying concepts, methods and possible tools used when undertaking VET policy evaluations.

Using VET policy evaluations as a platform for identifying and designing reform proposals in transition countries, as well as monitoring reform progress, the study is as much concerned with aspects of system analysis, as it is with elements of implementation research.

The study argues that there is no holy grail in terms of conceptualisation or methodology related to VET policy evaluations. The engineer's toolbox is of limited use. Similarly, management approaches based on refined analytical frameworks have been found to obscure rather than illuminate VET, including reliance on tools derived from system analysis. Thus, the only remedy seems to be the evaluator's broad understanding of the essential components of VET, of the relationships between them, of the fundamental logic between the system and its environment and, of change levers. This is an understanding that develops only through many years of apprenticeship and first-hand experience with VET policy evaluations, and not least through a close dialogue with key local actors.

Evaluators need to develop an understanding of both the starting points for reform in the given country context and strategic levers of change. The study emphasises that process evaluation, which acknowledges the role of organisational or administrative learning as part of an overall evaluation activity, is as important as impact studies. The study also argues that any VET policy evaluation needs to pay specific tribute, amongst others, to the professionalisation and continuous development of teachers and other education specialists, as they are key to the success of any systemic VET reform effort.

Presented in this study are the first results of a kind of ground-clearing work in VET policy evaluations. As such, we see our study as a basis for discussion, which opens a number of fields for further research and debate.

⁽¹⁾ In cooperation with Jean-Marc Castejon, Henrik Faudel, Barbara Kuta and Piero Benazzo.

Table of contents

1. Introduction	195
1.1. Methods of work	197
2. The limited transparency of concepts and methods used in VET policy evaluations	198
2.1. OECD	198
2.2. The building-blocks approach	198
3. Humanistic approaches to understanding VET systems	201
3.1. The hermeneutic approach	201
3.2. The structural-historical approach	201
3.3. Understanding and explanation	202
4. VET as a (sub)system and issues of internal and external consistency	204
4.1. System analysis	204
4.2. VET as a subsystem	205
4.3. Internal and external consistency	205
4.3.1. Internal consistency	207
4.3.1.1. Consistency between subsystems and clarification of pathways	207
4.3.1.2. Regional consistency	207
4.3.1.3. Consistency for the individual learner	207
4.3.2. External consistency	207
4.3.3. Regulating mechanisms	208
4.4. Examples of lack of internal and external consistency of VET reforms in Bosnia-Herzegovina, Estonia and Romania	209
4.4.1. Teacher training in Bosnia-Herzegovina (Durand-Drouhin and Bertrand, 1994)	209
4.4.1.1. Pedagogical Academy in Sarajevo (PAS) in charge of education of primary school teachers	209
4.4.1.2. Sarajevo University, Faculty of Philosophy, Pedagogical Department, in charge of pre-service VET teacher training	209
4.4.1.3. Pedagogical Institute, Canton Sarajevo, in charge of in-service VET teacher training	210
4.4.2. Curriculum reform in Estonia (Nielsen, 1999b)	210
4.4.3. National dissemination of pilot project results in Romania (Nielsen, 1999a)	211
4.5. Example of stressing system connections in a VET policy evaluation for Croatia	212
4.5.1. Decentralisation and school headmasters' and teachers' training	212
4.5.1.1. Governance	212
4.5.1.2. Finance	213
4.5.1.3. Curriculum (content)	213
4.5.2. Curriculum reform and teacher training	213
4.5.2.1. Design and development of new curricula	213
4.5.2.2. New equipment	213
4.5.2.3. Curriculum delivery	213
4.5.2.4. Teams of teachers	213
4.5.2.5. Curriculum assessment and revision	214

5. System change and the process and actors in VET policy evaluations	216
5.1. The complexity of systemic reforms	216
5.2. The path dependency of institutional change	217
5.3. Accomplishing change	217
5.4. The role of evaluations	218
6. Criteria, values and indicators underpinning VET policy evaluations	220
6.1. Evaluation criteria	220
6.2. Similarities and differences in VET reforms in western European countries	220
6.2.1. Improving progression from initial VET to higher education	221
6.2.2. Improving links between vocational education and employers	221
6.2.3. Improving the status of VET teachers and trainers	221
6.2.4. Improving the vocational curriculum	222
6.2.5. Summary	222
6.3. Potential and limits of the benchmarking approach	223
7. Conclusions and needs for further research	226
7.1. Conclusions	226
7.2. Needs for further research	229
List of abbreviations	231
Bibliography and references	232

List of figures

Figures

Figure 1: Basic concepts in evaluation	195
Figure 2: Denmark's ranking – education sector	224

1. Introduction

‘We shall not cease from exploration
And the end of all our exploring
Will be to arrive where we started
And know the place for the first time.’
T.S. Elliot, Little Gidding, V

National and international donor agencies have a shared understanding of how and why to undertake evaluations. Evaluation in its classical definition aims at comparing the design and implementation of projects or programmes to actual outcomes by analysing:

- (a) relevance to objectives and defined needs;
- (b) efficiency in providing inputs promptly and at lowest cost;
- (c) effectiveness in achieving planned outputs and immediate results;
- (d) impact on higher level objectives to which the result should contribute;
- (e) sustainability over time, and especially after

the project/programme inputs have been provided and/or external assistance stops (Grootings, 2000) ⁽²⁾.

The following scheme summarises the relations between these different criteria and aspects of projects/programmes:

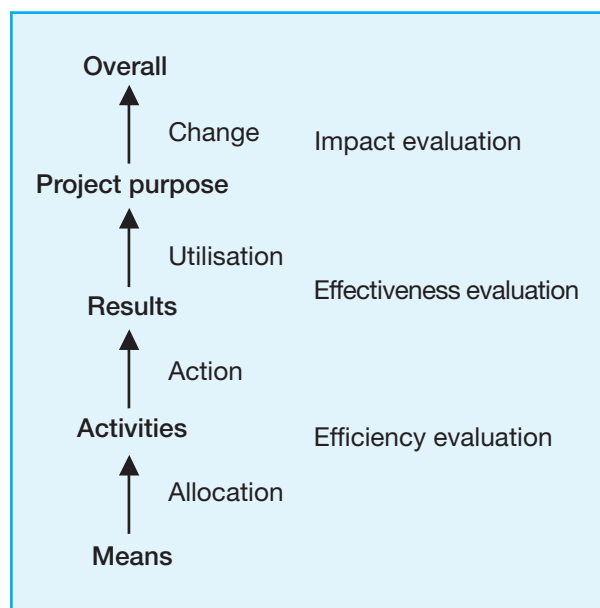
The scheme illustrates that:

- (a) efficiency evaluation has to do with the relation of means and expected results. It assumes that means and expected results are clearly formulated;
- (b) effectiveness evaluation deals with the relationship between achieved results and immediate objectives. Again, this assumes the existence of clearly formulated objectives;
- (c) impact evaluation concerns the relationship between immediate objectives and overall objectives. This assumes that both immediate objectives and overall objectives are clear (Grootings, 2000) ⁽³⁾.

However, evaluation increasingly becomes a continuous cycle that takes place at all stages of a reform or change process. The aim is to reflect on policies and practice with a view to influencing them. Such evaluation is meant to have an impact on decision-making processes, on how reform actions are designed or redesigned, on costs, communications, etc.

According to Windham (2000), there are ‘three primary purposes of all assessment work in the field of education policy: to assess the nature and magnitude of the opportunities and constraints that face the systems that provide education and training; to assist the government, the private sector and the individuals in establishing priorities within a resource-constrained environment; and to specify options for exploiting the identified opportunities and for dealing with the identified constraints. Alternative goals and strategies will be proposed from which government and its

Figure 1: **Basic concepts in evaluation**



⁽²⁾ Evaluation criteria and their definition are based on documents from the European Commission's Joint RELEX service for the management of Community aid to non-Member States (SCR – now *EuropeAid Cooperation Office*). However, identical presentations can be found with the World Bank, the Organisation for Economic Cooperation and Development (OECD), United Nations organisations and others.

⁽³⁾ Ibidem

partners can make concrete recommendations for policy reform.’

Evaluation in this sense is no longer just a reflection on the outcomes of some project or programme, but a type of consultancy work providing advice and guidance and the basis for informed decision-making. This makes criteria, such as relevance, effectiveness, efficiency, impact and sustainability relevant or irrelevant. However, project/programme evaluation and policy evaluation are often closely linked and feed each other.

Also, international literature suggests a trend away from evaluating the impact of single projects. A review of evaluation practice in education (OECD), VET (Grubb and Ryan, 1999) and labour market policy (Schütz et al., 1998, among others) signals a significant change in evaluation approaches towards target-oriented approaches in labour market policy and a systemic approach in VET. These shifts are caused by a growing awareness of the restricted contributions of single projects and the limited impact they can have, even if highly sustainable. Reviews of international cooperation (e.g. King and Buchert, 1999) similarly observe a shift away from project-based funding to policy assistance. As relevant publications suggest, the Danish international development assistance (Danida) and the UK Department for International Development (DFID) are increasingly following this approach, as is the European Commission in foreign aid policy *vis-à-vis* the countries and territories eligible for the MEDA programme (the Euro-Mediterranean partnership). However, while a pretty robust methodology for project or programme evaluation exists (even up to the point that there is now an evaluation of the evaluation methods), this is not the case for policy evaluation (Grootings, 2000).

We admit that the distinction between policy, which broadly encompasses political intention and may be expressed through several programmes, and particular programmes, which are themselves made up of many projects, may be somewhat arbitrary. Sometimes a policy is

virtually identical with a programme. However, the rationale for this distinction is to emphasise that we are more and more involved in broader policy evaluation. The evolving human capital imperative and its interrelationship with education encourages decision-makers to expand or alter their views regarding the nature of useful evaluation. Policy-makers increasingly want information regarding systemic, not only operating unit, performance. This is also the case for the partner countries with which the European Training Foundation (ETF) ⁽⁴⁾ is working. They ask for advice related to future policy directions rather than just support with the design of a certain project.

The ETF has both an analytical and a developmental role, i.e. directly or indirectly the Foundation is involved with the implementation outcomes of its evaluations. The main tasks include:

- (a) to identify gaps and needs for intervention or assistance;
- (b) to assist with the identification of needs or the formulation of policy objectives;
- (c) to identify and design reform proposals in transition countries;
- (d) to review progress and evaluate the outcomes of VET reform programmes;
- (e) to manage the dialogue with policy-makers and other stakeholders from our partner countries.

The analysis of an individual project or a limited number of themes would be too restrictive and the strategic perspective too narrow for the above assignments. Hence, VET policy evaluations are of strategic importance for the Foundation or, generally, for anyone who assumes, or prepares for, a development role in a domestic or a foreign context.

We explored the contribution of research and examined a number of examples from transition countries with a view to illuminating, where possible, VET policy evaluations. In line with the remit of the ETF, we have concentrated on, and used examples from, transition countries in central, eastern and south-eastern Europe, as well as from our Mediterranean partner countries.

The aim of our study is to clarify underlying concepts, methods and possible tools used in

⁽⁴⁾ The European Training Foundation is an agency of the European Union whose mandate is to support VET reform in its 40 partner countries, comprising central, east and south-east European countries, the successor States of the former USSR and Mongolia, as well as our Mediterranean partner countries and territories

VET policy evaluations. Our search for more conceptual and contextual clarity and possible instruments is, not least, connected with the hope that a shared understanding among evaluators can contribute to an increased coherence of policy recommendations and a clearer rationale for establishing priorities.

Chapter 2 points to the fact that there is rarely transparency for the methodologies deployed in (vocational) education (and training) policy evaluations, quoting the example of the Organisation for Economic Cooperation and Development (OECD). However, in a number of evaluations carried out or commissioned by the ETF, there is evidence of a building-blocks approach which was used to structure the evaluation and which is further explained in this chapter.

Given the limits of methods derived from management thinking, we argue in favour of humanistic approaches to understanding VET systems. Describing the hermeneutic approach, Chapter 3 discusses how a particular system can be understood. It also refers to the need to put certain phenomena in their historical context (the structural-historical approach) which helps to understand why existing practices, functions and structures are as they are. Finally, we refer to the continuous interaction between understanding and explanation, as explanations are needed when we fail to understand.

Chapter 4 examines VET as a (sub)system and elaborates on both the concepts of internal and external consistency of VET systems and the inherent tensions between flexibility and consistency, which an evaluator must grasp when undertaking VET policy evaluations.

Chapter 5 places VET policy evaluations in the context of system change. Viewing policy evaluations as a consciously organised, critical reflection and learning process involving the social actors and, hence, as formative and constructive, we emphasise their indispensability in any development process.

Chapter 6 explores the possibility of establishing predefined lists of evaluation criteria, values, policy priorities, indicators or other assumptions underpinning VET policy evaluations and argues why these should be subject to negotiations with the social actors involved. Furthermore, we review first experience with the use of the benchmarking approach on an international scale and point to its potential and limits.

Chapter 7 includes the executive summary and outlines needs for further research.

Finally, there is a list of references and a structured bibliography of background literature on the topic.

1.1. Methods of work

A working group was set up within the ETF to examine and compare concepts and methods used for VET policy evaluations. Group members reached understanding through collaboration and mutual learning. The study method was a structured collaborative writing process where working group members generated the results of their analysis in and through writing. Group members engaged both in an extensive literature review that included the European Commission, OECD, Cedefop, the World Bank, ILO, BiBB, the Leipzig and Vienna groups of transformation researchers, plus selected evaluation reports from our partner countries and many others.

The project also benefited from workshops involving both working group members and international experts. Two expert meetings were organised: one in February 2001 bringing together representatives from major international organisations operating in transition countries, and a second in November 2001 with evaluation experts from European countries. Moreover, all group members were involved, over the past two years, in a number of VET policy evaluations in various transition countries.

2. The limited transparency of concepts and methods used in VET policy evaluations

VET policy evaluations cover a multi-disciplinary, highly complex terrain. Concepts and methods are derived from sociology, economics, psychology, history, pedagogy, philosophy, management and organisational theory, comparative education and VET research, etc., whose scientific theoretical positions and methods have to be integrated. Furthermore, using VET policy evaluations as a platform for identifying, designing or redesigning and monitoring policies or programmes/projects, it was clear that some kind of implementation research would be as important, since, unavoidably, some form of intervention, or a type of applied, forward-looking research, is involved.

However, the methods and tools that support first analysis and then implementation are explored in international literature only in outline and often remain implicit rather than explicit. That literature conveys other aspects: the values and assumptions that underpin policy; the information and analysis that feeds in and out of policy formulation and determination; and the research base that exists or is created to support policy determination and implementation.

2.1. OECD

The OECD as one of the major international players in the field of (education) policy evaluation suggests that theoretical underpinnings derive from the choice of experts rather than being an explicit part of an overall approach. At the Foundation's workshop in November 2001, Whitman (2001) set out the approach of the OECD to system reviews: client country consensus; a framework of values and criteria; a careful choice of experts carrying their own cultural and research hinterland; and legitimacy from both the client country and the general secretariat of the OECD. The main purposes of the evaluations as such are:

- (a) to link educational policy to national, regional and economic issues;

- (b) to place each country's system in an international comparative perspective;
- (c) to make policy recommendations to government (being of greater benefit to partner countries and other donors by not itself representing a donor organisation).

While the two underlying themes (or values) are:

- (a) the emphasis on active labour market policies and their effectiveness;
- (b) improving labour force skills and competences through wide-ranging changes in education and training systems.

Whitman's exposition is useful, since elsewhere in OECD literature (even in country review outcomes) there is rarely transparency for the methodologies deployed.

2.2. The building-blocks approach

Analysing evaluations that had been carried out or commissioned by the ETF, we found that some form of the building-blocks approach had been applied in certain cases. The notion of building blocks was first mooted by Parkes (1995a). The approach was used for evaluations both in Bosnia-Herzegovina and in Moldova as a basis for constructing the VET 'green' and 'white' papers. Furthermore, the approach was used in the Maghreb countries (Algeria, Morocco, Tunisia) as a common framework for comparison, comprehension and subsequent cooperation; and in Uzbekistan as a means of evaluating policy determination and implementation. The approach has also appeared in project terms of reference drawn up by the ETF, most recently for Armenia, Montenegro and Turkey, etc.

The metaphor 'building blocks' referring to an effective VET (sub)system was introduced first to key actors from Baltic countries and then to a review of curricular reforms of the 1990s in nine central and east European countries. In their initial use, building blocks were defined in terms of functions or process. It was suggested that all successful VET systems, independent of their

cultural and historic contexts, need to possess the following ingredients (ETF, 1996a) (Parkes et al., 1999, p. 27):

- (a) 'to be able to define occupational sector priorities (on the best possible evidence available);
- (b) to be able to identify the appropriate occupational sector competences and skills required (and to construct the institutions and tools to do this);
- (c) to be able to turn these into curricular profiles and programmes and measurable standards;
- (d) to deliver these at school level (including the capacity to transfer from pilot to system level);
- (e) to help make the processes attractive to students and teachers (transferability, visibility and portability of qualifications for students and working conditions for teachers);
- (f) to provide for timely and effective feedback through evaluation, monitoring, quality control and tracer studies of school-leavers.'

It was recommended that these elements be undertaken in the context of establishing the balance between general and vocational provision and in the context of transparent and accepted approaches to standards, certification and qualification. They also have to be related to other factors: financing mechanisms; changes in the location of decision-making; a credible research base; and the development of management capacity, the acquisition of appropriate tools in management, as well as curriculum development *per se*.

While we can agree with the specified functions, which need to be undertaken in any modern VET system, there are some problems with the use of metaphors. It is, in essence, a functionalist analysis and should, hence, be called this way.

Added later on to the notion of function or process was the development of an architectural element. This was to offer an operational model for transition country working groups set up to analyse existing structures and practices and to make proposals for change in a consistent way, in order to have a common agenda among ministries and agencies. Eight topics (building blocks) were formulated:

- (a) educational management and administration;
- (b) curriculum, assessment and certification;

- (c) financing of VET;
- (d) the labour market and social partnership;
- (e) educational standards and quality control;
- (f) in-service teacher training;
- (g) legislation;
- (h) labour market and adult education.

Each of the building blocks was then analysed in detail, for example, by answering the following three key policy and action questions for the building block of legislation:

- (a) should there be separate or integrated legislation for general education and VET?
- (b) should existing legislation be left largely unchanged with modification to regulations only or should there be clear, new legislation, indeed an Education Reform Act?
- (c) if new legislation, should it be short, simple and transparent, leaving the detail to regulation?

Judging from the evaluation report of the respective Phare project under which the exercise was undertaken (Nielsen, 2001), Bosnia-Herzegovina can be considered a very successful example, in which the building-blocks approach was used to structure and systematise the analyses by the actors involved and then to provide a framework for a concept paper for legislation. The building-blocks approach helped establish a common conceptual grasp of the issues at stake and a common language in which a relatively large group of key actors could discuss structures, functions and institutions of a VET system in transformation. According to the report, it turned out to be a useful tool for specifying well-grounded and specified VET reform strategies, encompassing all elements of VET systems.

In the cases we reviewed, the building-blocks approach was an attempt to provide a simple, transparent vehicle for managing the dialogue among country key actors, donor representatives and consultants, i.e. between the researcher and the practitioner. It was an attempt to balance actual complexity with simplicity and transparency in the use of tools, which is consistent with Grubb and Ryan (1999) who recognise the need to find a way to express a holistic approach in a manner convincing to policy-makers: 'In the selective use of evidence, policy makers have avoided the complexities of theory and methodology. Instead they have often used social science results in simplified ways, to tell almost

commonsensical stories consistent with the particular ways in which they want to frame problems. In this habit they are reinforced by journalists, whose methods of personalising issues by concentrating on the experience of individuals –

story telling – makes the issues more vivid in the public mind, but [...].

However, as the ending indicates, the approach also leaves a number of unresolved questions which will be discussed in the following chapters.

3. Humanistic approaches to understanding VET systems

Instead of methods derived from management thinking, we agreed on a phenomenological approach, in which the emphasis is on describing phenomena, as they appear to the observer, instead of attempting to guess or theorise about what might be behind these phenomena. The basis is first and foremost to establish a genuine understanding of what you observe.

Having said this, we asked ourselves whether we can identify a number of methods or tools (*Krücken für die Gedanken*) actually made use of by experienced VET policy evaluators. Which methods do the latter use to come to an understanding of VET? And what is the correct relationship in evaluations between understanding and explanation?

3.1. The hermeneutic approach

In more specific terms, we may talk of a hermeneutic approach that covers a number of stipulations and various methods related to the understanding and interpretation of phenomena. Hermeneutics play an important role in human and social sciences; an important thinker in this field is Hans-Georg Gadamer (1972).

A hermeneutic approach seeks to elucidate what happens when you try to come to grips with a text or any other phenomenon in society. In the description of how to come to an understanding, a specific concept is normally used: the concept of the hermeneutic circle. This refers to the fact that the particular element is understood in terms of the totality, while the totality is, at the same time, understood through its constituent parts. Through our interpretation work we will achieve a more and more secure, detailed and varied understanding by way of a continued circular alternation between studying parts (system elements) and totalities (the total system). There is an additional circular process in the interpretation work: the interaction between the preconception (*Vor-Verstehen*) of the phenomenon, which the researcher/evaluator brings with

him/her, and the new understandings worked out through the process.

An important point for discussion on hermeneutic approaches is the question of whether objectivity is possible. Some hermeneutics would argue so. Karl Popper, for example, claims that you can avoid wild interpretations and arbitrary postulations, which are a clear risk in hermeneutic understanding, by going through a number of well-conceived loops for testing your interpretation hypotheses. However, Gadamer questions this position.

Gadamer (1972) insists that it is not possible to come to an understanding of a phenomenon, for instance a foreign VET system, just by following certain methodical procedures. According to Gadamer, the interpreter cannot abstract him/herself from history and culture in trying to understand, as the researcher is her/himself part of history and culture. Understanding always takes place between two entities both placed in their own historical and cultural contexts. What the researcher/evaluator must do is openly recognise and clearly articulate this tension. Another conclusion for us is the involvement of evaluators coming from the reviewed country itself or, in the context of external peer reviews, from countries with similar education systems.

3.2. The structural-historical approach

Building blocks or ingredients may be a useful, simple tool to structure (or manage) evaluation, but the approach also leaves a number of unresolved questions. One is the use of metaphors instead of clearer concepts; but let us continue metaphorically. If these (eight) building blocks are the bricks, what then is the mortar holding the building together? Can we provide a conceptual framework for grasping not only the parts (building blocks) of the VET system, but also its historic roots and dynamic relationships? Which architecture structures the VET system under

observation and gives the constituent parts their specific meaning? Are there different systemic logics behind the way building blocks are put together in the different countries? How can we dig deeper into the inner driving forces behind the configuration of the building erected over many years?

A structural-historical approach (*Strukturell-genetischer Zugang*) could be applied. Here, the analysis would start with an overview of existing practices, functions and structures as they appear to the (experienced) evaluator. Essential structures are then traced back to their origin (genesis), and an understanding/explanation of the historical context is reached in which the specific phenomenon was born or established. Described through the comparative, structural-functionalist mirror, for instance, the overall Danish VET system may seem a wonderfully harmonic building that consists of highly refined building blocks (Cort, 2002). But analysed historically (through the crystal ball instead of the mirror), this VET system is a patchwork of political compromises reflecting what was achievable under given historical-cultural circumstances over the last 100 years. How do we come to such an understanding of other VET systems in evaluations, using the building-blocks approach?

Another question related to the understanding of the systemic logic behind the building blocks is the challenge of change. How do VET systems change and what are the driving forces behind change processes? Even if we get the description and understanding of the building blocks right in VET policy evaluations, we cannot at all be sure to have fully established the capacity to predict change or to formulate the right intervention strategies to achieve change. Is there a strategic lever at force in VET system change? What is (empirically and theoretically) known about dynamic forces, catalysts for change, etc., and what is the right balance between top-down (politics) and bottom-up (market) approaches? What follows is that the building-blocks approach must be accompanied by a deeper insight into organisational and institutional processes of change. We will come back to these issues in Chapter 5.

Yet another problem is how to cope with the risk of preconceptions and nationally biased understandings of what makes up a good VET system. It is not uncommon to see the foreign

evaluator measure the distance between VET reform initiatives in a transformation society against the advanced state of his/her own system. The first question is: are the building blocks *per se* culturally bound? The next question is: how can we make use of them in evaluations, being nevertheless reasonably objective and transparent? One way around this dilemma is to make statements of values behind evaluations explicit, when using the building-blocks method. Chapter 6 will deal specifically with this issue.

3.3. Understanding and explanation

It has been argued above that the phenomenological/hermeneutic method is primarily used to achieve an understanding rather than for explanations. If we accept this point of departure, then we would argue that we should not give up using the building blocks as an interpretative framework between the evaluator and the VET system under observation, but we should definitely put more emphasis on explanations.

Understanding and explanation are different ways of (re)cognition. Understanding is a more immediate experience or recognition of a phenomenon. Sometimes, if a phenomenon is difficult to understand, we need explanations to come fully to grips with, say, a component of a VET system and its precise role within the system. When you explain something, you establish some distance to the phenomenon under observation. On the basis of explanations, you may often understand the phenomenon better or view it from a different perspective. To explain something is to state causes behind the phenomenon under analysis, external to the phenomenon itself. To understand a phenomenon is to give a reason for it, which is internal; meaningful understanding presupposes a communicative community between the evaluator and what is being evaluated. Understanding a phenomenon, *in casu* the VET system, means that one can supposedly see (recognise, realise) how the elements fit together, and see the meaning of the phenomenon (intention, purpose, function), often in an immediate, even intuitive way.

The relationships between understanding and explanation are tricky; one form of recognition is

not better or more correct than the other. The argument here is simply that we should reflect on these connections and deliberately seek to establish a continuous interaction between understanding and explanation.

The following are but a few examples of useful explanatory frameworks:

- (a) economic or labour market or sociological laws and other forms of determination of the concrete conditions under which a phenomenon is active, for example, the conditioning factors behind VET systems, such as demography, trends in labour market developments and broader educational policies;
- (b) functionalist explanations referring to the totality in which the phenomenon to be explained is placed and a description of the precise function it serves; for example the role of a national VET system for societal functions such as qualifying, socialising, sorting, and being a depository of young people;
- (c) historical-cultural explanations behind the occurrence of different institutions and practices in contemporary VET systems, for example: why is there such an assortment of independent VET providers (production schools, etc.) in Denmark compared to Sweden? Because of *Grundtvig* and the free school and folk high school tradition;
- (d) structural(ist) explanations of phenomena which can only be uncovered through an in-depth analysis, such as the specific connections between the economic, social and political systems in social structures, making transition of VET systems in many former communist countries very difficult;
- (e) system-analytical explanations showing how changes in one component will have spill-over effects on other system components or on the total system, cf. the following discussion on internal and external consistency, such as the role played by access to higher non-university education as a determinant for the attraction of VET.

Central to this reasoning is that understanding and explanation feed each other, that understanding is the point of departure for VET policy evaluations, and that explanations are needed when we fail to understand.

But we easily and often fail to understand. Awareness of one's own limited or culturally bound

understanding is as important as ever. Four examples may illustrate the problems evaluators face:

- (a) the trendy catchwords broadcasted by the OECD, the European Commission, the World Bank, etc., in recent years have mainly focused on modern competences, such as adaptability, changeability, responsibility for one's own learning, etc.;
- (b) the cultural school of thought in European VET research (around Anja Heikkinen from Jyväskylä University, Finland) argues that this is a frontal attack against VET institutions which the labour movement has been able to establish only after more than 100 years of hard struggle. The change of emphasis from the sociological qualification to the psychological competence is bound to remove social control from VET. And it undermines the social category of the skilled worker;
- (c) the lifelong learning concept is now recommended everywhere. As evaluators we tend to measure VET system developments against the indicators found in or around the European Commission's memorandum on lifelong learning (European Commission, 2000a). We tend to do so even in countries where the most rudimentary conditions for initial VET are not in place;
- (d) in our thinking on modern and adequate VET provision, we tend to promote output-based, modularised and credit accumulation-based models with little emphasis on process factors (curricula, teachers, pedagogics, time-servicing requirements). We promote the Anglo-Saxon models without remembering why they had no choice but to configure the system the way it is;
- (e) the complete destruction of the East German polytechnic education system was the result of a lack of understanding of the intrinsic qualities of that system. Entirely lost is the expertise developed in work-based learning principles (*Lernen in der Arbeit*), didactics and practices, which were further developed in the socialist countries at the time than anywhere else.

Therefore, besides developing further the building-blocks approach by including explanations, and in order to avoid hobbyhorse explanations, we have to cultivate, individually and in a community of VET evaluation practitioners, a new attitude towards argumentation. We need a more open discussion practice: a discourse.

4. VET as a (sub)system and issues of internal and external consistency

In our effort to understand VET systems (or the system context of VET policies), it seems obvious that we have to apply some kind of system analysis. However, it is quite difficult to define a VET system:

- (a) how is the system universe to be delineated?
- (b) what are the constituent components, parts, units?
- (c) which relationships exist between units?
- (d) what matrix of dominance patterns the interplay of units?
- (e) where are the boundaries of the system located?
- (f) which relationships (metabolism) exist between the system and its environment?
- (g) what is the prime mover or the strategic lever for change?

The set of questions is very complex and forms part of a broader scientific context. It is helpful to go back to the roots of system analysis and to analyse, from a theoretical standpoint, issues related to internal consistency and external consistency of VET systems and their consequences for VET reforms, and hence VET policy evaluations.

4.1. System analysis

The VET field is short of a comprehensive theory and in many ways remains a very young research field. Perhaps one could classify the research tradition as follows. First, there is work within the tradition of general educational theory. A second group of topics relates to the study of those organs of the VET system which exist by law and are recognised in official documents about VET structure, functions and actors. Third are studies which take these formal institutions as given and explore the forces, including the informal institu-

tions, which set them in motion and link their activities.

A whole range of topics is opened for discussion, as soon as one asks the question: what makes the VET system tick? One can then explain things in terms of functions; we can look for systemic interrelations between the data, with a view to discover what goes with what, what changes in parts produce what reactions in the whole and in other parts. We can treat functions as heuristic, as a notion that will help us to spot relationships and so to define structure as a system of parts with internal and external relations. An interesting example of this approach was the early 1990s research component of the European Union's (EU) PETRA programme on innovation of VET programmes. The research question asked was 'through which mechanisms are VET programmes renewed?' instead of asking 'who does what when?'.

General systems theory is a field of thought that has been growing in strength over more than 50 years. Its origins go back to biologists who opposed reductionism, i.e. an analysis of living things merely as a sum of parts (which was a view contradicted by observation). Parallel movements existed in the psychology of perception and learning (Köhler, 1959) ⁽⁵⁾, and in social anthropology, in which structure/function analysis already began to flourish in the 1920s. But the systemic approach is more than just an example of the use of biological analogies to describe society. Much impetus was given by mathematicians and electrical engineers during World War II and immediately after.

General systems theory would not claim to be empirical; rather it claims to be axiomatic, in that the mathematics conceived first for the electronic systems has been so far generalised that it is applicable to any systems whatever, provided that the rules are followed: systems are isomorphic.

⁽⁵⁾ Interestingly, the *Gestalt* concept has been applied very fruitfully by German VET researchers; cf. Felix Rauner, University of Bremen/Institut Technik und Bildung (ITB), and Gerald Heidegger, Universität Flensburg/ Berufsbildungsinstitut Arbeit und Technik (BIAT), who make use of the *Prinzip Gestaltung* as a way to conceptualise innovation in VET.

The meaning of the term ‘system’ is often confused. The most general definition is formulated by the founding father of the general system theory, von Bertalanffy (1950), after he had noticed that in physics, biology, psychology and social sciences it was no longer acceptable ‘to explain phenomena by reducing them to an interplay of elementary units which could be investigated independently’. ‘A system is a set of units with relationships among them; the word “set” implies that the units have common properties – the state of each unit is constrained by, conditioned by, or dependent on the state of other units. The units are coupled. Moreover, the system as a whole has “got something” which its components separately have not got. Systems may be concrete or abstract, systems may be “open” or “closed”, and systems can be analysed on the principle of systems and sub-systems – systems within systems within [...] ultimately a wholly general system. There are theoretical difficulties about this ultimate system: but the problems of supra-systems and sub-systems, levels and boundaries, etc., are generally manageable in a practical way.’

The German sociologist/philosopher Niklas Luhmann (2001) has built up a general system theory of society, where the educational system is a subsystem having its own logic, laws of motion, discourse, etc., and is even further differentiated.

4.2. VET as a subsystem

VET as a subsystem can be seen like other living organisms as an open system in a steady state; it depends on self-regulating mechanisms to maintain its boundaries and its continued existence within these boundaries (meaning primarily boundaries which distinguish the VET subsystem from other (sub)systems of the same society). The steady state depends on a balance of inputs and outputs. The inputs are demands and support: support makes the system strong enough to process demands and to produce outputs in the form of qualifications. The VET system sits in an environment – the total social

system – and there is a continuous feedback into the VET subsystem, signalling whether outputs produce good or adverse effects in the environment. A systemic approach to VET evaluations focuses on the analysis of relationships, communication channels, and responsiveness and adaptability, based on the fundamental understanding that changes in one component lead to changes in other components and in the system as a whole.

It is important to understand the specific logic of the VET system, to explore whether the system is internally and externally consistent, and in which way the system needs to be regulated to balance available tensions.

4.3. Internal and external consistency ⁽⁶⁾

Improving VET is possible only through a systemic approach, and we provide arguments below why this is so. Evaluations would have to establish to what extent VET policies have followed a systemic approach, applying in particular the concepts of internal and external consistency. At the same time, the evaluator needs to be aware of important system-inherent tensions that are not at all easy to reconcile.

In transformation societies the main driver for reforms is the fact that the whole logic of VET systems has radically changed. Before the collapse of the planned economies, there was a close relationship between the human resource needs of companies and vocational schools. Based on a precise manpower forecasting (planning) model and a vocational guidance system whose main aim was to steer young people into certain occupations in line with available job positions, schools could almost exactly cater for the needs identified by companies. People tended to remain in their jobs for a lifetime. With the change to market economy principles, this system no longer worked. The labour market underwent dramatic changes within a short period, calling for different skills than those catered for by schools. Schools had been slow to adapt, results

⁽⁶⁾ Section 4.3 and subsections 4.3.1–4.3.3 draw heavily on Durand-Drouhin and Bertrand, 1994.

include a large number of young unemployed and inadequate skills upgrading or retraining system for adults. There are tremendous conceptual, political, financial, structural, organisational and content-related problems to be solved at the same time. This concerns internal factors of VET, but above all the need to establish new mediating mechanisms between education and work to replace the former administrative ones.

Most countries have placed new emphasis on the need for more responsiveness of VET systems to changing needs for education and training. Two key concepts, flexibility and consistency, are associated, to variable degrees and in variable combinations, with the search for responsiveness.

Responsiveness is needed with regard to:

- (a) constantly changing and largely unpredictable requirements of labour markets for skilled labour;
- (b) growing demands by young people for higher levels of education and for more diversified and individualised learning processes and pathways.

Responsiveness can be interpreted and may be achieved in a variety of ways. In institutional and management terms, three major developments can be identified.

First, there is a clear trend towards more decentralised governance and management of education systems in general and of VET in particular. Examples include France, which in the early 1980s, delegated authority and responsibility for the location and equipment of vocational schools largely to the regional level, as well as several Scandinavian countries which transferred responsibility to local authorities and schools. The latter were given a great deal of initiative, including curriculum design and implementation, in order to satisfy diverse needs in local economies and of individual students.

Second, the role of central government in educational planning is declining, while industry is called upon to participate more actively in the design and provision of VET and of related certification and qualification systems. While the participation of employers and their organisations is expected to contribute to improved responsiveness of VET to both quantitative and qualitative demands in the labour market, the involvement of trade union organisations and workers' councils varies from country to country.

A third trend is the development of training in enterprises and in private institutions and the related emergence of a training market where public and private institutions may be competing. This trend is related to the growing importance of adult training, which can itself be seen largely as a response to the demand for a more flexible workforce.

In terms of the structure, content and organisation of VET, the search for responsiveness has in particular led to:

- (a) the diversification of training programmes and pathways;
 - (b) the broadening of curricula;
 - (b) the increased duration of initial training;
- and in some countries:
- (a) the modularisation of courses and certificates;
 - (b) the introduction of competence-based learning and assessment.

All these developments are assumed to contribute to the adaptability of individual workers and to the flexibility of education systems. However, they also raise serious questions about the types of qualifications which young people acquire, about the costs and effectiveness of education pathways and about the consistency of the rules of the game prevailing in different sectors of national education systems and in the labour market.

These issues seem to be of interest to central and east European countries, which are going through a process of transition, from centralised and rigid education and training systems towards other types of systems, expected to meet the demands of market economies. Considering that, in the present context of uncertainty, these demands are to a large extent unpredictable, the need for adaptability and flexibility would seem to be even more pressing in countries in transition than in most west European countries.

The implications of such orientations concerning the structure, organisation and management of VET systems should be carefully examined. There are certain tensions between objectives of flexibility in VET on the one hand and consistency on the other. The underlying question is whether flexibility and consistency can be conceived of as complementary rather than contradictory features of education and employment systems and which regulatory mechanisms deserve particular attention in this respect.

This question will be considered from three points of view: the internal flexibility and consistency of education and training systems; the external responsiveness of education and training systems and the consistency between their internal organisation and objectives and the functioning of labour markets; and, finally, different approaches to system regulation.

4.3.1. Internal consistency

4.3.1.1. *Consistency between subsystems and clarification of pathways*

Introducing flexibility by diversifying education pathways has also meant the creation of ever more sinuous, long and unfocused educational careers. Moreover, while countries have moved closer to flexible and open systems, their strategies of reform have tended to concentrate on one type and/or level of education and training at a time, rather than explicitly aiming at complementary transformations of all stages and in all relevant areas. And yet the consistency of learning opportunities and pathways across types and levels of education and training is essential.

One important aspect of the internal consistency issue is the relative attractiveness of VET *vis-à-vis* academic education. Many efforts to attract young people into vocational streams do not take a sufficiently broad view of the relationship between the different streams. For instance, one of the reasons for the preference for academic streams is the wider opportunities that they provide for further studies. But attempts to open up access to higher education to vocational students often fail, because universities operate on a completely different logic and use different selection criteria.

The two examples illustrate that improving VET is possible only through a systemic approach, encompassing all the components of the system. Such an approach may conflict with a policy which would promote a high degree of decentralisation and of privatisation.

4.3.1.2. *Regional consistency*

When far more autonomy is given to schools, not only from the point of view of management but also of teaching content, the problem of possible imbalances and inequalities between regions, schools and individual students might arise. For the central and east European countries, with

their strong egalitarian tradition, this raises the question of how to overcome such problems.

Most English-speaking countries, with traditionally highly decentralised systems, shared a concern about the multiplicity of assessment and certification systems. This was one of the reasons why the UK and other English-speaking countries engaged in the development of national vocational qualifications. They are not, in contrast to many continental European countries, related to remuneration systems through collective bargaining at national or sectoral levels. This is considered as an additional element of flexibility. However, the value of such qualifications and their motivating effect on young people and adult learners may also be weaker than that of more holistic and rigid qualification systems, such as exist in Germany.

4.3.1.3. *Consistency for the individual learner*

In countries like the UK, the definition of national vocational qualifications is associated with a modular approach to training, allowing students to choose their own personal range and sequence of course/modules (which can be provided by different institutions) and to accumulate corresponding certification units. Young people are thus – to some extent – encouraged to define their own individual learning pathways. Modularisation is expected to allow for more rapid responses to changing labour demand. Flexibility is pursued both from the point of view of students' interests and of labour market requirements.

One of the possible drawbacks of modular approaches is the risk that young people will acquire only loose collections or incomplete puzzles of narrowly defined skills or competences. In such systems, responsibilities for the level of skills and the relevance and completeness of skill profiles with which young people leave the education system is largely left to young people whose information, strategic planning capacities and financial possibilities vary greatly.

4.3.2. External consistency

External consistency refers to the relationships between educational objectives and the economic and social context. In reality, there are potential problems, such as the fact that the employment system is not always able to articulate its needs and that there are tensions between

the pedagogical logic of school systems and competence needs of employment. Furthermore, the drive to maintain a holistic approach of training for an occupation, and to develop national qualification systems as a means to increase internal and individual learner consistency, may conflict with deregulated labour markets and the short-term recruitment and training practices of companies. It is also realised that education cannot by itself solve the problems of the employment system. See, for example, Hodgson (2001), Paquet (2001), Miegel and Nölke (1996), as well as Grubb and Ryan (1999): 'Education without suitable employment, and specific skill training without jobs requiring such skills may be valuable in their own right but they cannot enhance economic conditions. And so the other conditions necessary for education and training to be effective – the employment necessary, the capital required, the institutions that can give these arrangements some permanence – also need to be carefully understood, and the most successful programmes carefully consider the nature of local employment.'

In evaluations commissioned by the ETF we have found a number of examples of internal and external consistency lacking in the VET policies implemented by transition countries of central and south-eastern Europe. We include some of them in Section 4.4. In contrast, in an evaluation in Croatia, we have tried to highlight system connections which need to be considered when designing a consistent (in the given case, EU-funded school headmasters' and teachers' training) programme (Section 4.5).

4.3.3. Regulating mechanisms

The foregoing observations underline the need for regulation of VET systems themselves and of the interface with their social and economic environment in order to find the narrow path between two dangers. On the one hand 'a centralised, uniform and rigid system would not meet the needs of a modern economy, would limit adaptability and threaten the ability to innovate' (Wolf, 1993). On the other hand, leaving individual schools and training institutions without any common framework of objectives and organisational conditions would not satisfy the long-term economic goals of quality and effectiveness and is likely to increase geographic imbalances and social inequalities. Ideally, successful system

regulation requires both clear policy orientations and effective regulation mechanisms, which need to be reviewed by VET policy evaluations. Clear policy orientations are particularly important in a context of strongly decentralised systems. As far as possible, policies should be based on both a sufficient capacity of technical analysis and a high degree of social dialogue. In institutional terms, this implies the need for:

- (a) 'a framework for consultation among the various actors, which, at the national, local and/or sectoral level, guarantees some degree of continuity, coherence and consistency, especially between education systems and labour markets;
- (b) some kind of research and development structure, providing information and technical support to decision-makers; [...] this includes [...] the development and monitoring of labour force information systems, the development of learning theory and appropriate teaching and training methods, etc.;
- (c) clearly defined and agreed financing agreements which are a major element of the system regulation. This raises policy – or political – questions about the respective role of governments, enterprises and individuals; [...]
- (d) finally, there is the broader issue of evaluation. The more decentralised the system, the more necessary it is to set up mechanisms allowing for the evaluation of its output and effectiveness. Certification is an important element of evaluation and it has to be seen, not only from the point of view of recognition on the labour market. The above comments concerning the limitations of a market-driven approach do not deny the fact that the ultimate value of qualifications depends on the market and not on administrative decisions. But the need for consistency pleads for some degree of cooperation between the actors and for some form of institutionalisation in order to regulate the market and transcend its most immediate and specific demands.' (Durand-Drouhin and Bertrand, 1994).

This brief overview of requirements or conditions of successful regulation refers to an ideal situation. In reality, it is clear that such requirements cannot easily be fulfilled, especially in the present context of transition countries. To

mention only a few points, the institutions and the identity of the various actors are only emerging and the inadequacy of technical instruments is reinforced by the lack of resources. But, in one way or the other, responses to these problems, adapted to national cultures and traditions will have to be found.

4.4. Examples of lack of internal and external consistency of VET reforms in Bosnia-Herzegovina, Estonia and Romania

To understand the importance of systemic consistency (and to which phenomena the concepts refer) when evaluating VET policies, examples shall be provided from three transition countries. More examples are available from the ETF.

4.4.1. Teacher training in Bosnia-Herzegovina (Durand-Drouhin and Bertrand, 1994)

Part of the questions during the evaluation were addressed to actors and institutions outside the EU Phare VET reform programme with a view to assessing whether framework conditions would be conducive to the longer-term sustainability of the pilot project. Obviously, one central factor here is vocational teacher training, which had not originally been a project component. The pilot schools had been asked only to specify teacher-training needs in relation to newly introduced curricula. This was changed later, and a budget reserve of EUR 200 000 allocated to run a programme for teacher and management training.

The following questions were asked in three institutions visited: were existing teacher training institutions involved? And were they familiar with the curriculum principles applied in the Phare VET reform programme?

4.4.1.1. Pedagogical Academy in Sarajevo (PAS) in charge of education of primary school teachers

The PAS department for Chemistry was visited. The reason for choosing chemistry was that in almost all pilot VET schools visited, it was argued that pupils are below standard and that particularly the new Phare module in chemistry is very problematic and probably the most difficult

subject. The two assistant professors interviewed had never heard of the Phare VET programme and its renewed module in chemistry, neither were they aware of poor end results of pupils in primary schools.

The conclusion is that a clear problem exists here in terms of the vertical integration of the education system, which has a negative impact on Phare programme results. Only the Ministry of Education can solve this problem by promoting internal consistency measures.

4.4.1.2. Sarajevo University, Faculty of Philosophy, Pedagogical Department, in charge of pre-service VET teacher training

An interview was carried out with the Dean, Prof. Dr Vlado Sucic. Pre-service vocational teacher training is subject-based and covers the full range of subjects taught in secondary schools. Teacher training is carried out according to a consecutive model: first, students graduate in a subject (e.g. engineer), then there is a six-month pedagogical-psychological-didactical-methodical training course, which includes school practice. After two years of teaching practice at a VET school, the candidate takes the final state examination and becomes a certified teacher. This model is also offered for teachers of practical subjects.

Until now, the University's Pedagogical Department had always been involved in the design of new curricula for all subjects taught. This was done to ensure that new subject matter and innovative methods were integrated immediately into teacher training. However, the department was not involved in the Phare programme and did not even know of its existence.

The fact that no inputs from changes in VET equipment, curricula and methodology are fed forward and channelled directly into pre-service teacher training is a serious problem. The development model is not optimal: too little emphasis is placed on the teacher training system and the training of teacher educators. In addition, VET teacher training would capitalise on becoming an important research field in Bosnia-Herzegovina. Any support programme should focus on the teacher educators to ensure a multiplying and lasting effect and to ensure that activities are 'anchored up' within existing providers of pre-service (and in-service) teacher training.

4.4.1.3. *Pedagogical Institute, Canton Sarajevo, in charge of in-service VET teacher training*

Tasks of the pedagogical institutes in Bosnia-Herzegovina in general include:

- (a) advising ministries of education on curricula and new legislation;
- (b) collecting statistical and other information from schools;
- (c) evaluating the quality of schools;
- (d) advising on the appointment of teachers and assessing and confirming their qualifications;
- (e) providing professional development opportunities for teachers and setting up in-service training courses;
- (f) evaluating school development plans.

The institute has traditionally promoted a subject-based approach to curriculum design and delivery, which led to some opposition with respect to the newly introduced module-based curricula in Phare pilot schools. An interview was held with Mr Vreto Sadik, inspector in the field of mechanical engineering at VET schools. His specific responsibility covers subject matter and training practice in these schools. He is familiar with the Phare programme and had participated in two seminars in the previous year, one on curriculum development and teacher training, the other on the dissemination on the Green Paper. He formed part of the team who approved the new equipment specifications. But he is only one of very few.

Pedagogical institutes need to be involved in development projects along the lines of the Phare VET programme. The latter focused on 12 pilot schools, while the pedagogical institutes were totally forgotten. Subject inspectors did not receive any training; they see their subjects disappear in the long run or being parcelled out into new modules, and they are less and less able to perform their supposed roles when going to the schools. This is a serious barrier to broad acceptance of the reform. In fact, it may give ground to a large group of key actors who veto reforms.

The above three examples demonstrate a sub-optimal reform environment resulting from a lack of internal consistency of reform measures introduced.

4.4.2. **Curriculum reform in Estonia (Nielsen, 1999b)**

The EU-funded Phare 1995 VET Reform programme in Estonia aimed at assisting the country in kick-starting VET reform, which would

be key to developing a workforce required for a modern market-driven economy. The specific activities to be undertaken embraced the following main components:

- (a) curriculum development;
- (b) teacher training;
- (c) provision of equipment;
- (d) establishment of links with partner schools in European Union countries;
- (e) support to the design of a VET reform strategy and dissemination of results.

The curriculum design principle chosen in Estonia is a flexible, modular-based and employer-led model. In the development of curricula, Estonia follows a model similar to the Irish (FAS) one, but adapted to the Estonian context. 13 selected pilot schools developed modularised curricula to nationally approved industry standards. Social partners, in particular the Chamber of Commerce and Industry, are actively involved, at national level, in the definition and approval of occupational profiles and, at local level, through school-based curriculum design groups developing course modules to national standards.

The system logic of VET has changed. A methodology has been developed and successfully piloted to identify skill needs and educational goals and to translate occupational profiles into curricula. This has functioned well. Employers and their representatives have started to influence VET provision by articulating their skill requirements and requiring schools to produce skilled workers according to these needs. The dynamics of the VET system have been increased and self-regulating mechanisms built into the very core VET.

The results achieved show that the adopted strategy served as a catalyst for building the thrust and the momentum needed in the new approach to curriculum development. The basic institutional frameworks are in place.

However, the curriculum support infrastructure is extremely fragile. Human resources in the VET Department of the Ministry of Education are severely limited. Some years ago the Ministry had already transferred its responsibilities for curriculum innovation to the national centre for examination and qualifications (NCEQ). The centre, currently engaged in other assignments such as handling national examinations, is methodologi-

cally and technically clearly too weak. It has not been in a position to develop sufficient capacities to absorb Phare programme results, such as endorsing the newly developed modules or driving the reform process forward. In 1999, about 130 modular descriptors that had been submitted to the centre were still awaiting their approval and accreditation. The centre, in its function as the national vocational qualification institution in the Anglo-Saxon philosophy, constitutes the basic cornerstone in the whole VET system logic. If the system is to work well, the capacities of the NCEQ to set up a national qualification structure and to act as the national award and accreditation body will have to be built up quickly. Staff needs to be recruited and trained.

This is a clear threat to sustaining VET reform results. The NCEQ needs to be given a clear mandate and to attract new experienced staff. In addition, it requires immediate technical support with a view to building up its capacity. Simply not endorsing the modular descriptors developed under the Phare programme, as was the case, puts both the momentum of the whole curriculum innovation process and the still fragile interaction between the employment and the education system seriously at risk.

It follows that there is a lack of external consistency, owing to problems of consistency between educational objectives on the one hand and the economic and social context on the other. Although attempts to increase the responsiveness of VET to the changing needs of the economy were a stated priority, there was only limited evidence of the implementation of this objective, and the educational infrastructure has continued to operate primarily according to its own logic.

4.4.3. National dissemination of pilot project results in Romania (Nielsen, 1999a)

The Phare VET Reform project in Romania selected 75 VET schools to act as pilot schools, some 10 % of all schools. At the end of the reform programme the country was faced with the challenge of generalisation of results. The government then decided that the new curricula were obligatory in all schools, without any new equipment or any proper training. The national inspectorate, whose actors were themselves only half-familiar with the new curriculum principles, was supposed to support this vast implementation effort.

The decision to generalise the pilot school curricula indicated an intention to take the Phare VET reform methodologies from the experimental to the systemic level. However, it is difficult to see how this could be achieved without a considerable investment in equipment and training of both teachers and school managers, in short strengthening material as well as human resources. Pilot schools and teachers are not in a position to generate their own resources for this huge effort. The necessary critical mass of change agents may be in place to consolidate what has been achieved. The teachers and curriculum development officers trained under the Phare programme could have played a crucial role in training the staff of the majority of vocational schools not involved in the Phare programme. However, a broad nation-wide implementation effort requires careful planning and substantial resources.

In all types of pilot projects there is a risk of encapsulation. As a rule, the best institutions are selected and, during the piloting phase, they are granted good development conditions, i.e. they are freed from the difficulties that exist in the ordinary structure. When it comes to the transfer and broad systemic implementation of pilot results, all the real world problems and barriers suddenly turn up again. The risk of barriers in the ordinary structure is always a serious factor to be reckoned with when going from the phase of 'VET reform (pilot) project' to the phase of 'national VET reform strategy'.

One other impeding factor in the model school approach is the fact that a number of elitist schools are given all the equipment, all the coaching, all the study tours, all the development assignments, etc., leaving the initially backward schools another couple of years behind. It is not clear how Romania (or any other country) will cope with this problem.

Another threat to the wider dissemination of curricular reforms, as initiated under the Phare programme, is the relatively low level of teacher qualifications. A major national in-service teacher training programme is required to make VET reforms happen at school level. Modularisation, as the basic pedagogical principle, implies a radical change of working conditions for Romanian vocational school teachers. In many ways, the teaching of modules could be seen as reduced qualification from the point of view of the professional teacher. So what is needed is not only technical up-skilling

but also awareness-promoting initiatives on a broad scale.

The reform of VET systems is more than a single act establishing a new legal framework. A change in the logic of the system can only be realised gradually by those who are involved in VET themselves, especially at the local level and in the schools. Since there are many people involved, a tremendous learning process is necessary, which needs to be facilitated and intensified through proper intervention and guidance.

After a long period of central regulation and bureaucratic administration, it is particularly important to depart from the traditional top-down conception of VET reforms and promote self-responsibility and self-initiative for local actors. But such bottom-up strategies, that give the initiative for educational change to the individual schools, may easily lead to fragmentation and competition between schools, as a consequence of the pilot-school based reform strategy.

Romania shares this challenge with other central and east European countries. It will be necessary to integrate the various changes introduced in a more coherent change strategy. From a strategic point of view, one should reflect once again on the interrelationships between the various aspects and issues and ask the fundamental question: where should one start in order to trigger a development process that will lead to the desired outcomes? In particular, it is important to explore whether the applied strategy of bottom-up curriculum-modernisation-through-pilot-schools, which has been forced upon the country by the EU Phare programme, has had the intended impact. And how, given the experience so far, is it possible to transcend from the level of piloting to the level of systemic reform. A fundamental discussion about the role of the vocational teacher and trainer may now well be of a more strategic value than the continuation of curricular reform, which has dominated VET reform so far.

We conclude that the challenge of national dissemination has caused tensions between the objective of flexibility in VET on the one hand and consistency on the other. This also raises the question of different approaches to regulatory mechanisms with a view to ensuring internal and external consistency.

4.5. Example of stressing system connections in a VET policy evaluation for Croatia ⁽⁷⁾

While the examples given in Section 4.4. indicate a lack of consistency in one form or another, a further example shall show how emphasis has been placed on system connections in a VET policy evaluation for Croatia. The original assignment given to the ETF was to analyse the training of VET teachers and trainers in Croatia and, on this basis, to come up with proposals for an EU-funded development programme.

Basic problems identified in Croatia include:

- (a) the non-existence of a genuine and holistic pre-service VET teacher training system (would-be VET teachers read a technical subject plus general pedagogy; there are no in-company training practices);
- (b) the in-school training year of would-be teachers is not well supervised;
- (c) the central Institute for Educational Development offers, through a centrally published catalogue, a list of teacher training courses established centrally on the basis of perceived needs for teacher training and available capacities;
- (d) in-service teacher and school headmaster training schemes are highly limited in general; training is neither linked to school development nor to teacher performance evaluation or career progression.

Based on the fact that changes to one part of the system have an impact on other parts of the system, the challenge for us was to design a reform process that builds upon the internal links and develops both the framework structures and the actors at the same time. We highlighted the following system connections.

4.5.1. Decentralisation and school headmasters' and teachers' training

4.5.1.1. Governance

School headmasters, administrators and teachers will have to share the burden for a number of new educational management issues. This implies a radical change, as schools must be managed in a

⁽⁷⁾ Viertel, 2003. The text draws on Nielsen, 2002.

different way. Apart from pedagogical innovation, schools will have to establish a strategic planning and budget execution capacity. Both school headmasters and selected teachers will be involved at the strategic level of the new school organisation. This calls for training to develop the necessary skills.

4.5.1.2. Finance

Money will (partly) be transferred in the form of block grants, and schools will have to strengthen their capacity for planning and economising in the context of budgetary restrictions. They may also have to look for new income sources and will have to form local partnerships with parents, education authorities, employment services and employers. This will have immediate consequences for school headmasters, administrators and teachers alike: all of them will have to learn how to act as entrepreneurs.

4.5.1.3. Curriculum (content)

A modern VET system is based on a framework rather than detailed regulations, to give some freedom of action to the institutions providing the education and training services. Curriculum development and delivery will, therefore, be decentralised to some extent. At the central level, both clear objectives in terms of learning achievements (competences) and frameworks for content will be formulated, while schools will have to determine, to a higher degree, both the more specific content elements and how learning objectives are to be achieved (learning methods). As emphasised also in both the White Paper and the Ministry's concept paper, modern learning theory calls for a shift from teaching to learning and from a teacher to a learner-centred focus. In this context, headmasters will play a key role in promoting new pedagogical approaches in their schools, and VET teachers and trainers have to be equipped with the necessary skills to master these new assignments.

We are aware that the needs for training in relation to the change processes are enormous, and that training should start as soon as possible. Without such training and school development processes being initiated on a massive scale, curriculum change and VET modernisation will not be accomplished. As this is also a never-ending process, resources have to be freed up in the budget for this task.

4.5.2. Curriculum reform and teacher training

A decentralised VET system ascribes new roles and responsibilities to all involved. New curriculum principles will involve all teachers at a vocational school. Curriculum-led VET reform has considerable implications, including the following.

4.5.2.1. Design and development of new curricula

Within given objectives and frameworks defined at national level, teachers will be actively involved in curriculum development: they will have to translate overall learning goals into specific subject matter, and determine the actual teaching plans and the methods of teaching/learning. In addition, selected teachers will also have greater responsibilities in other fields as outlined above.

4.5.2.2. New equipment

New curricula imply the use of new or updated learning (pedagogical) equipment. The modernisation of school workshops calls, first, for teachers and instructors to master the new machines themselves and, second, for the ability to make good pedagogical use of it, i.e. to create meaningful learning opportunities for students. An intensive training in vocational didactics is necessary. The (practice-related, partly in-company) programme for the training of teachers and trainers within the dual VET system, which was implemented by the Chamber of Crafts and Trades in collaboration with western partners, may serve as a good example in this respect. Although we are aware that it is impossible to send all teachers and trainers abroad for a certain period, learning on the spot from the best vocational schools or companies is probably the most effective way to learn about the new, often tacit knowledge and skills required.

4.5.2.3. Curriculum delivery

New student-activating methods must be introduced, such as problem-solving organised as group work, project work, etc. The focus is on training teachers in developing the capacity to organise new learning processes, to stimulate the learning motivation of students and to establish better learning environments. This would have to go hand-in-hand with the development of new teaching material.

4.5.2.4. Teams of teachers

The work of teachers and instructors will follow new principles of organisation. The need to

develop key occupational competences on the basis of a much better integration of general subjects, subjects of vocational theory and subjects of vocational practice requires a much closer cooperation between all teachers and trainers in a vocational school. We have witnessed an excellent illustration of this new approach in a private economics gymnasium, where various subject teachers work together to cluster a number of formerly separate, but related subjects into broader subject fields. The new approach presents a radical challenge vis-à-vis current school practices: the teacher becomes part of a collective planning body composed of several teacher colleagues who altogether feel responsible for the holistic (more broadly occupation-oriented rather than narrowly subject-oriented) competence development of an individual student. To achieve this goal, a lot of school-based teacher and trainer training and development will be required in the coming years, as an inherent feature of true curriculum reform. Again, the school headmaster has to understand these processes to be able to provide direction and leadership, orchestrating the efforts of the teacher teams.

4.5.2.5. Curriculum assessment and revision

In a decentralised VET system leaving substantial freeway to the schools, mechanisms will nevertheless have to be introduced to assure quality to the level of agreed national standards and a purposeful and efficient use of public funds. As a result of pressures from the financing authorities, but also a number of civil and economic actors to which the school becomes increasingly accountable (students, parents, employers, etc.), vocational schools will have to reflect on the range of their programmes and the quality of their services. The best schools in Europe have developed a genuine evaluation culture as a shared preoccupation by all school staff, which encourages feedback from their external clients with a view to improving learning processes and outcomes continuously. Such an evaluation culture can only be established on the basis of mutual trust and must reflect the values of the school, which can vary considerably from school to school. Once more, school headmasters would have to play a pivotal role in establishing such a culture.

As not all the necessary changes in the Croatian VET system could be fully implemented immediately, a careful choice had to be made of elements that are critical to reform. Our choice of

priorities was based on the following criteria: instead of a series of separate interventions in the fields of pre-service and in-service VET teacher training, we recommended trying out a coherent, systemic approach based on partnerships between schools, universities, teacher training institutions and companies.

Furthermore, our approach focused on the principles of learning on the job or action learning, meaning that learning should be as closely related to the changing day-to-day work processes as possible. This way, theoretical knowledge would be combined with experience-based learning as part of the change process and become applied knowledge or action competence (*Handlungskompetenz*). Rather than taking teachers out on external courses (as is the current practice in Croatia), learning outcomes are optimised when teachers and school headmasters learn through first-hand experience, being supervised by teacher-mentors in their original work environments. Having gone through such a process themselves, teachers are more likely to practice the same new style of learning with their students in day-to-day teaching practice afterwards. Moreover, involving all school staff in a collective learning process, rather than training school headmasters and teachers individually, seemed the most promising way forward towards school development. Headmasters and teachers need to share the same vision and see a role for themselves in a future reformed system. Also teacher trade unions needed to be taken aboard.

The activities which were finally recommended for inclusion in the EU-funded development programme were, amongst others:

- (a) the development of change agent team strategy and the training of a sufficient number of change agents as a way to introduce new VET reform concepts and train teachers at school;
- (b) the development of the mentor as one of the profiles of education specialists, the selection of excellent teachers specialising in certain fields and their training as mentors;
- (c) the design of a training programme for school headmasters which should be owned and multiplied later on by the School for School Headmasters in Croatia;
- (d) the design of new methods for analysing training needs, the pilot-testing of training courses delivered in VET schools and the implementation of a system to monitor the

- relevance and effectiveness of training before, during and after the courses;
- (e) the implementation of a number of school development projects as the principal approach to training;
- (f) the identification of a central teacher training institution that would accompany the development process, and own and disseminate further all innovations introduced into the system.

5. System change and the process and actors in VET policy evaluations

Apart from the systemic approach, the evaluator needs to develop an understanding of how systems change in the interactions between actors, that innovations need to be taken forward not only to implementation but also to institutionalisation, and what are possible strategic levers for change.

5.1. The complexity of systemic reforms

To illustrate the complexity of systemic reforms in VET and the role of evaluations in them, let us first have a closer look at how they are defined. While, currently, there is no agreed definition of systemic reforms, Grootings (1993) describes them in the context of the following post-socialist stages of reform in transition countries (the first stage was added later by Bîrzea, 1997):

- (a) corrective reforms that are initiated with immediate repairing objectives in countries emerging from a war or deep economic regression period;
- (b) modernising reforms that are interventions aimed at reducing gaps and catching up with western institutions. They are especially active at the level of curricula, teaching and learning methods, examinations, and school textbooks;
- (c) structural reforms that are targeted at the structures, legal framework and management of educational systems;
- (d) systemic reforms that are deeper and have a global character because they call for a genuine change of paradigm in terms of educational policy. They are aimed not only at the curricula or the legislative framework, but at the internal logic of education and its relationships with the global social system. A systemic reform examines the key elements of every educational policy: the role of the state, relations with the labour market, the financing system, efficiency control, the normative role of national standards, etc.

A study by the US Department of Education (1995) offers the following definitions for systemic reforms in education:

- (a) 'systemic reform addresses all of the mutually reinforcing structures, processes, and activities within the educational system, recognising that altering any one part of the system necessarily impacts on all other parts;
- (b) systemic reform requires system coherence through the integration of policy and practice;
- (c) systemic reform constitutes a mainstream activity of all organisations involved, not an alternative or special programme;
- (d) systemic reform requires strategies that help develop and mobilise the conceptions, skills, and motivation in the minds and hearts of scores of educators;
- (e) systemic reform requires the development of routine mechanisms for bringing people together across roles, within and across organisations, for developing and maintaining shared direction and understanding; and to maintain strong communication among all of the constituent parts of the system;
- (f) systemic reform in education addresses the preparation, continuing learning and working conditions of [...] educators in all roles – teachers, principals, counsellors, specialists, paraprofessionals, central office and higher education personnel.'

What is interesting to note in the above definitions is that:

- (a) the professionalisation and further development, in a structured and continuous way, of teachers and other education specialists are considered key to the success of any systemic reform effort;
- (b) higher education plays a central role in implementing reforms of the VET system. Much will therefore depend on new school-university partnerships.

In other words, changing the VET system – as any social system – is a complex, multi-layered process, comprising all system levels:

- (a) the primary learning process;
- (b) the organisational level;
- (c) the institutional frame;
- (d) the policy and legal frame.

5.2. The path dependency of institutional change

The organisational settings and arrangements, and related social values and behavioural patterns, have grown historically as the result of long, non-linear social development processes. According to Grootings (2002a), ‘institutions result from social relations among people and imply therefore power and trust. They are embedded in historical contexts and part of wider institutional frameworks in which different institutions are related to each other’. Taking the example of curriculum reforms, Parkes et al. (1999) illustrate their institutional embeddedness as follows:

‘Curriculum is perceived as a balanced relation between the potential and interests of individuals and the requirements of society. From the individual point of view “it” can be seen as the totality of measures, interactions and experiences within an organised learning process. However, individuals can make up their curricula only within patterns laid down by organised bodies and social institutions within society. These institutional frameworks are anchored in the specific social system of a society. They offer opportunities to the individual who in turn is constrained by requirements and standards to become a “responsible” member of society, in whatever way this is defined. The educational field has its own bodies and institutions to develop a framework to respond to the temporary and long-term needs of society, as well as the wishes of the individual to promote his/her personal development. ... The main actions, actors and institutions involved in the development of a curriculum are part of “institution-building” involving developing “institutional homes” for key functions and activities.’

Grootings (2002b) argues that there is an institutional logic that is country-dependent, and reform through institutional change is determined by the institutional context of the country. They represent what was possible to achieve at a certain point in time rather than being a perfect making. They are unique for each country, which explains the uniqueness of reform designs for each country. Similarly, current OECD economic thinking (the development of social capital) suggests that reform, and economic development, is best undertaken through the mutation of historic institutional structures.

5.3. Accomplishing change

It is also important for the evaluator to understand the ways in which understanding and support by the actors necessary for implementation can be gained, as a lack of this has often led to failures of reform projects. Reforms will have to be carried out by existing staff, and despite widespread agreement with the global policy objectives and the recommended improved policies, there may still be great discomfort felt by the reform actors at the prospect of changing traditional ways. Accomplishing change is about reversing deeply embedded policies and strongly held beliefs. It is about preventing resistance to change or attenuating its impact.

The prevailing political culture will determine the accentuation of approaches to reform – with top-down and bottom-up reform approaches at the two ends of the spectrum. Again, there is no wrong or right in favour of the one or other approach. Our experience tells that contrasting organisational forms are required for interventions at the central level and the level of organisations, respectively. A high degree of bureaucracy is likely to make structural reform easier, while non-bureaucratic reforms are needed for curricular innovation and learning.

While, in transition countries, system change was traditionally designed by the centre and decreed from the top, the political culture is now changing. Local agents claim a stake in the design of reform concepts and, as we have seen, resistance may build up at an early stage when Governments, or foreign consultants, present ready-made concepts rather than involving local change agents in reform design from the very outset.

According to Paquet (2001), the defining characteristics of democratic change include bottom-up governance schemes and a new form of transversal coordination, while collaboration becomes the new categorical imperative. The centre will focus on norms and the periphery on delivery; for the two to function together, a clear sense of public purpose is needed, as well as new partnerships and skills. The agencies need to be granted the necessary powers to organise activities and become negotiating arenas, providing space for interaction, the (re-)definition of activities, as well as monitoring. A shift is

taking place in the policy management of public services: the stress is no longer on inputs but on outputs. The government defines objectives and sets frameworks while allowing a flexible style of decentralised management that is capable of rapidly adapting to new demands. This in turn calls for a decentralised structure, and for new forms of horizontal accountability, for the system of governance to be effective.

Such a process of cultural transformation and the translation of policy into practice is almost always an extremely lengthy process. McLeish argues in her introduction to *Process of transition in education systems* (McLeish and Phillips, 1998) that the completion of the transition process at the structural-legislative level in no way implies that educational transition at the micro level has been achieved. To change a label is easy, to effect a comprehensive change in practice is very difficult. Peter Drucker (1999), whose speciality, since the 1950s, has been to distil current thinking and research for the practical manager at the level of the organisation, makes the following remarks in his last book: 'There is an enormous amount of work to be done in organisation theory and organisation. There are only organisations which have distinct strengths, limitations and specific applications. A given organisation fits certain tasks in certain conditions and at certain times [...] One cannot manage change; one can only be ahead of it [...] Change is necessarily a series of small incremental changes.'

We conclude that system change will have to build on the given historically grown institutional structures. It is likely to be achieved only through small, incremental change in narrow and targeted areas and only where there is equilibrium between radical change and traditional forces. Change requires, according to Paquet (2001, see above) a clear sense of public purpose, new partnerships and new skills, as well as careful policy coordination, compensatory mechanisms and collaboration in adequate forums for consultation and decision making.

5.4. The role of evaluations

Planning system change would be the art (rather than the science) of designing a multi-purpose, multi-actor, multi-stage process comprising

coherent actions across many interdependent areas. This process must be based on a sound understanding of the institutional logic of the system and a number of hypotheses about the reasons for lack of internal and/or external consistency and how they could be overcome. Due to the multitude of factors impeding a system that is in a steady state of development, and the unpredictability of human behaviour, such planning models and their outcomes must always remain imprecise. OECD (2000a) argues that 'the outcomes of educational processes are multi-dimensional, underlying factors remain imperfectly defined.'

Being aware of the imperfection of any planning model, it becomes essential to build into the change process, strategic phases where social actors come together and reflect and learn with the ultimate aim of verifying or rectifying hypotheses. This explains the need for evaluations not as one-off events, but as a continuous cycle.

Evaluations take on different roles in such a change process. They can be commissioned by the (national or international) funders of reform projects with a view to providing feedback on progress and establishing whether they have met their targets. In this case, evaluations would be summative (evaluation of results/impacts); they would serve as accountability tools and be largely limited to the projects in question. However, as already argued in the introductory chapter, project evaluations provide too limited a scope for assessing the complex nature of changes in VET within their social/ institutional contexts. That is why also 'summative' evaluations need to be complemented by formative elements. What formative literally means is having influence in forming or developing.

Evaluations can be organised as a conscious, critical reflection process at various stages of the reform process. They involve the reform actors, arrange for learning and negotiating opportunities and help shape or reshape policies. As such, they are part of the development process. They are constructivistic in that they deny that there is an objective knowledge of the world (the VET system) and in that they put actors at the centre of the evaluation rather than its outcomes.

Stern (2003) describes the following steps of such an evaluation:

Steps of a policy evaluation process

- (a) identification of stakeholders;
- (b) holding a series of initial discussions;
- (c) feedback;
- (d) identification of areas of agreement or disagreement;
- (e) use of other sources which could inform stakeholders;
- (f) reaching the best possible consensus.

Consensus creates one priority. A second need is capacity building by means of a dialogue, in simple and transparent terms, with the key actors so that it is possible for them to internalise and own the outcomes of evaluation. This will, subsequently, allow them to have the understanding and political will to move to strategic implementation both at macro and micro levels.

Inspiration and guidelines can be found in the so-called empowerment evaluation school in US university evaluation research. But this approach is also well known in Europe. For example, the UK Tavistock Institute of Human Relations (TIHR) makes use of an approach described as 'development evaluation' (Stern, 1989). It involves a number of elements:

- (a) it offers participants in the setting being evaluated a voice in shaping the evaluation agenda;
- (b) it involves an active feedback policy;
- (c) it sets a clear focus on utilisation and implementation.

The approach is based on clusters of shared professional values, one of which is the commitment to action research.

Nyhan (1998) outlines the lack of a research tradition in VET (as opposed to general education). He reviews the approach of Kuhn (1996) in rooting rationality in particular political and cultural contexts, and Carr and Kemmis (1986) in accepting that the social actors involved in transformation need to be engaged in a kind of action research. In doing so, he points the Leonardo da Vinci Programme firmly in the direction of action research: 'how can we change things at present to ensure a better future?'

Gisela Shaw (1999) further elaborates the Nyhan position: 'This type of research (action research) is based on two fundamental assump-

tions. The first is that social research, including research in VET, is ill-advised, if it tries to copy the positivistic model of the natural sciences, relying primarily on quantitative and empirical methods and aiming at research results of universal validity allowing scientific predictions; instead, social research ought to accept that human actions and behaviour can only be interpreted and illuminated and that such interpretation and illumination requires taking into consideration the social context in which they occur (Hans-Georg Gadamer's hermeneutic/interpretative approach; Gadamer, 1972). The second goes beyond this position by aligning itself with a claim embedded in Jürgen Habermas' critical theory: human science being rooted in the emancipatory human knowledge interest (rather than in the technical or practical knowledge interest) cannot stop short at interpreting and illuminating human actions and behaviour but ultimately aims to equip individuals with an understanding of how best to pursue their objectives in a rational way (Habermas, 1972). Although acknowledging the problem inherent in the notion of action research, i.e. a deliberate blurring of the boundary separating the researcher from the practitioner, Nyhan nevertheless highlights the fruitful implementation of action research in educational and training programmes in a number of countries [...]'.

The action research strategy offers channels for continuous dialogue with the relevant stakeholders (policy-makers, actors) about the validity of implementation of the concepts being studied, throughout the project: 'An action-research project attempts to build concurrent dissemination activities into the research process' (Nyhan, 1998, p. 28).

This means that the process of undertaking an evaluation – i.e. the way in which it is negotiated and managed and how it impacts on the subjects of an evaluation – is itself seen as a legitimate object of study. It follows that there is a thin line between evaluations, or action research, and quite specific consultant activity.

It follows also that the experience of the evaluators is key (as the OECD also suggests – see Section 2.1). Long apprenticeships seem to be necessary, whereby the junior evaluator would try to learn from the senior evaluator, thus going through the Dreifuss & Dreifuss five-stage process from beginner to expert.

6. Criteria, values and indicators underpinning VET policy evaluations

6.1. Evaluation criteria

In the introduction to our paper we have mentioned evaluation criteria, such as relevance, effectiveness, efficiency, impact and sustainability, which are important to project evaluations. It is clear that, for the purpose of policy evaluations, we would have to alter or redefine the set of evaluation criteria.

The OECD (Windham, 2000), for example, has used the following evaluation criteria for reviews of national education policies/systems in south-east European countries:

- (a) appropriate political foundation and support for the reform process;
- (b) equity in access, attainment and achievement;
- (c) flexibility in planning and implementation;
- (d) support for an interactive planning model involving cooperation among national, local, and institutional levels;
- (e) incorporation of intersectoral and intrasectoral coordination;
- (f) development of appropriate regional, national and international emphases;
- (g) affordability;
- (h) sustainability;
- (i) efficiency;
- (j) encouragement of supplementary resource mobilisation.

To evaluate reform policies meant to aid systemic change, as defined in Section 5.1, the following criteria may be useful:

- (a) the political foundation and support for the reform process;
- (b) the visions guiding the reforms;
- (c) the leadership driving them;
- (d) the structures and resources for planning and implementation, including the knowledge and research foundation upon which they are built;
- (e) the opportunities for learning needed to sustain them;
- (f) the mechanisms for communication used for intrasectoral and intersectoral coordination;
- (g) the organisational arrangements designed to support them;

- (h) the strategies used to implement them;
- (i) the structures and mechanisms used to monitor and evaluate pilot implementation, and take corrective action, where necessary;
- (j) the affordability/economic viability of interventions, as well as the structures and mechanisms used with a view to including innovative results in mainstream provision.

However, as illustrated above, a universal manual is difficult to assemble; a list would depend on the scope and purpose of the evaluation exercise. We suggest that, following the development evaluation approach, these evaluation criteria were subject to negotiations with the actors concerned prior to the start of the evaluation exercise.

6.2. Similarities and differences in VET reforms in western European countries

The same applies to the question of underlying themes and values. Evaluators often come with preconceived notions: VET policies in transition countries tend to be evaluated against western standards and practices. Can we, indeed, identify some common policy priorities (and indicators) which, sooner or later, lead to some convergence of systems and which could, hence, facilitate VET policy evaluations in transition countries?

Over the past decade, the VET systems of many west European countries have come under reform pressures as a result of persistent or new problems, such as:

- (a) the academic drift;
- (b) the poor motivation of students on VET programmes;
- (c) the declining investment of time and money of employers in VET (even in countries with strong traditions of employer involvement, such as Denmark, Germany and Austria);
- (d) global competitive pressures which are putting strains on the traditional social partnership approaches;

- (e) vocational teacher training (polarised between German university studies obligatory for all teachers and the zero solution in the UK);
- (f) the lack of parity of esteem between general education and VET;
- (g) the failure to provide future employees with new types of knowledge and skills which they will need in a fast changing global economy;
- (h) the problem of organising learning environments and learning processes which stimulate context-free learning of generic skills; etc.

Lasonen and Manning (2001) have shown, in an interesting typology, that west European countries have actually applied very different strategies to address these problems. They are summarised as follows.

6.2.1. Improving progression from initial VET to higher education

There are two distinct approaches to improving progression from initial VET to higher education; the access approach and the diversification approach. Denmark, Austria and Finland, on the one hand, have emphasised the diversification of higher education and the upgrading of existing vocational higher education by creating a new tier of vocational higher education institutions, where programmes have been extended in length. Many countries, such as Denmark, Finland and Norway have combined the diversification of higher education with creating opportunities for students on initial vocational courses to progress to higher education by obtaining dual qualifications.

England and Scotland, on the other hand, have concentrated on improving access to a single unified university system. However, a clear hierarchy remains between the old or traditional universities, which recruit almost entirely from students with academic qualifications, and the new universities which are more open to students with vocational qualifications. Local provision of further education colleges, often in partnership with universities has been expanded.

Most countries want the proportion of students in higher education to be expanded to at least 50 % of each cohort. In unified stratified systems, as found in England and Scotland, nearly all higher education institutions have equivalent status as universities. In divided systems, such as those emerging in Denmark, Austria and Finland, and already found in Germany, some higher

education institutions are specifically designed to have a vocational role and not be universities.

The general conclusion is that the question of parity of esteem of VET needs to focus on post-compulsory education as a whole and not just the provision for 16-19 year olds. As the numbers opting for general education at upper secondary level continue to rise, the parity of esteem issue shifts towards higher education.

6.2.2. Improving links between vocational education and employers

Major differences between the countries include:

- (a) those with a dual system of apprenticeship based on social partnership (Denmark, Germany and Austria) in which employers have a direct role in decision making about VET, which is based initially on their recruitment of apprentices;
- (b) the state-led systems of the Nordic countries in which governments have actively involved employers in VET curriculum issues through consultative committees at local and national levels;
- (c) the more voluntarist systems, such as England and Scotland and, to a lesser extent, Spain, where governments have attempted to involve employers through work-based training schemes, and through employers' membership in governing bodies of schools or colleges and regional training councils.

Voluntarist systems rely on creating the conditions for new types of learning partnerships between employers and educational institutions. The problem is how to increase the involvement of small and medium enterprises with little spare staff capacity for supporting partnerships.

A variety of ways for ensuring greater employer involvement in VET is being explored. In Finland, France, Norway and Spain a greater regional emphasis is allowing individual institutions to develop closer links with local employers. Work experience inherent in curricula is seen, especially in Finland, France and Spain, as an important driver for establishing new links with employers and enhancing the future employability of young people.

6.2.3. Improving the status of VET teachers and trainers

Improving VET is intrinsically linked with the quality of teachers. Various patterns are emerging

in European countries. First, a much greater emphasis is being placed on improving the quality of vocational teachers rather than workplace instructors; only in countries with a dual system are serious steps being taken to upgrade the qualifications of trainers. In England and Scotland, the focus on workplace trainers is limited to their (students' performance) assessment role. Second, general subject teachers in vocational schools are typically better qualified than vocational subject teachers (the dual system-based countries are, again, exceptions here). Most countries are trying to standardise qualifications and limit traditional regional differences (as in Spain).

Finland is the only country with a statutory in-service training requirement (five days a year). Most progress in terms of structural parity has been achieved in the three countries with a dual system (Denmark, Germany and Austria) and more recently in Norway, where the qualifications for VET and general subject teachers are equivalent and programmes are shared.

In England and Scotland, staff development and initial training of vocational teachers is left to individual institutions. Despite an increase in the numbers of qualified VET teachers, formal qualifications are still not a requirement for vocational subject teachers in colleges, though they do have to pass a test if they want to assess student performance rather than just teach certain types of vocational courses. Steps are being taken to establish national standards. One important step was not to introduce fees for the training of (would-be) vocational (or further) education teachers, placing them in an equal position with general subject teachers at primary and secondary level.

6.2.4. Improving the vocational curriculum

School-based systems, such as those in France and the Nordic countries, have increased the general education component of vocational courses, as well as the choices available to students. In Finland, this increase is expressed through the concept of integrated learning, whereby:

- (a) schools can decide to increase the general education content from one fifth to one third;
- (b) students can compose their personal study plans;
- (c) all programmes include work placements;

- (d) the vocational content has been strengthened by integrating it with activities from working life.

Mixed systems which are becoming more linked or unified, such as in England and Scotland, give priority to generic skills such as numeracy, literacy and team working, which employers claim are important, and to the mixing of general and vocational knowledge.

Systems with strong apprenticeship traditions, such as in Germany, try to enhance the vocational knowledge component of VET programmes through work process knowledge. Denmark is introducing greater flexibility into the dual system though the overarching concept of pathways. The Danish pathways approach stresses the support of student choices through the development of individual course plans within a modular curriculum and the need to complement vocational specialisation with broader-based studies. Vocational curricula in Spain seem to be developing against the European trend with a reduction in the amount of general education, shorter courses (more attractive to students and employers) and, in line with other European countries, more opportunities for local partnerships between schools and employers.

6.2.5. Summary

What the above analysis illustrates is that many of the west European countries have not tackled some of the fundamental problems in VET themselves; so there is no blueprint for a good model. EU experience has shown that neither the common factors pressing on national policy formation for education and training, nor the priorities that most EU Member States share, have led to a uniform pattern of convergence (even if some common trans-European trends have emerged). We are aware that the non-existence of clear-cut policy priorities or indicators may create a dilemma for the objective evaluator. A particular donor may impose them, but more easily on the evaluator than the specific government concerned. A possible way around the dilemma of preconceptions or nationally biased understandings is to start an evaluation by presenting the building blocks and formulating ideal practices against which existing practices would then be evaluated. Finally, the analysis shows that adopted strategies for VET system renewal or reform reflect the different educational

histories of each country rather than the problems these strategies are designed to address. They cannot be detached from their institutional contexts, as we have also argued in Section 5.2.

This is confirmed by Young (2001) who analyses why some countries have so far remained partly immune from the pressures to develop outcome-based qualification frameworks of the kind found in the United Kingdom and other anglophone countries. This is so despite the need to rethink the role of qualifications in light of global economic changes and the related changes in skill and knowledge demands (Lasonen, 1996) (Lasonen and Young, 1998) (Young, 2001). He shows that there are built-in contradictions with EU countries having quite divergent philosophies and practices, making in particular a distinction between an outcomes-based approach (English-speaking countries) and an institution or process-based approach (most notably those associated with the Germanic and Nordic traditions of education and training).

6.3. Potential and limits of the benchmarking approach

Despite the diversity of systems at EU level, attempts have recently been made by individual countries and at EU level to apply a benchmarking approach, first, to employment policy and subsequently to VET policy. What are the defining characteristics of this tool, what are initial results, and does benchmarking generally point the way forward in VET policy evaluations?

There are numerous definitions of benchmarking, the simplest being that benchmarking is the process of learning by making comparisons. Although simple, it captures some of its essential features: namely, that it is deliberate learning process and that it involves comparison. From the early days of benchmarking, companies have used it as a way to compare their own processes, practices and performance with other similar companies with a view to making improvements. More specifically, benchmarking in its classic form is a process of studying, adopting and adapting successful processes and practices

from other organisations to improve processes, products, and services in one's own organisation.

Since then benchmarking has become an umbrella concept used in modern management of both private and public organisations. The increasing use of benchmarking has also meant that some of its basic features have, in some cases, been dropped, which makes it difficult to distinguish it from comparison or comparative analysis.

To talk about benchmarking as a meaningful activity we need to view it as a formal and deliberate process of comparison, aiming at detecting weakness and creating ideas for improvement. Benchmarking is normally distinguished from purely analytical methods of comparison in that specific performance indicators must be developed and units must be found which can be considered as best performers concerning the chosen indicators.

The comparison then aims at finding options to make progress towards the position of the best performer (Grootings, 2000). Essential features include:

- (a) the need for improvement;
- (b) identification of a benchmark, i.e. agreeing on a benchmark that can serve as a reference point for undertaking the comparisons;
- (c) the translation of the analysis into concrete steps for improvement.

For several years the Danish Ministry of Finance (1999) has undertaken benchmarking exercises for various sectors to assess where Denmark stands compared to other countries with respect to a number of issues that determine prosperity and welfare ⁽⁸⁾. A best practice approach was followed by choosing a fixed group of countries consisting of France, Germany, Great Britain, Japan, the Netherlands, Sweden and the US who are considered to be among the world leaders in one or more relevant fields. Major areas of prosperity and welfare were identified, and each area compared, using a number of indicators. A ranking between the two best or the two worst scoring countries for each indicator was judged as being above or below average. In the area of education the indicators below were used, resulting in a ranking for Denmark which is shown on the left.

⁽⁸⁾ Similar evaluations have been done for Finland: *Benchmarking Finland – an evaluation of Finland's competitive strengths and weaknesses* and Sweden: *Quality comparisons in education and the employment offices*.

Figure 2: **Denmark's ranking – education sector**

Above average	• Number of hours of teaching per teacher in basic vocational education • Pupil/teacher ratio in primary and secondary education • Small proportion of young people who have been unemployed for at least one year
Average	• Total expenditure per pupil in the educational system • Annual number of hours of teaching in primary and lower secondary school – 12-year-olds • Proportion of population with upper secondary education • Proportion of population holding tertiary education qualification • Maths skills • Natural science skills – upper secondary education • Reading skills of 14-year-olds • Relative unemployment for 20 to 24-year-olds
Below average	• Number of hours' teaching per teacher in primary and secondary (general and vocational) education • Annual number of hours' teaching in primary and lower secondary schools – 14-year-olds • Natural science skills after 8 years' schooling • Reading skills of nine-year-olds
Source: Danish Ministry of Finance, 1999.	

As the Danish example illustrates, benchmarking is used solely to compare educational systems on the basis of their results, whereas a qualitative analysis of the practices, processes and methods that have led to certain results is not made. In the given example, no rationale was offered for why, for example, the reading skills of nine-year-old students are poor. Also, it is taken for granted that scoring high is good, while good reading skills are no guarantee that they could be readily applied to learning in everyday life. That is why the whole debate about which indicators to

choose for a benchmarking process is so important and needs to be considerably advanced ⁽⁹⁾.

To us it is doubtful whether the Danish benchmarking exercise was more than just an international comparison of selected educational indicators. It should be mentioned that the Danish Ministry of Finance clearly recognises that softer educational objectives, which are valued highly in the Danish context, are not sufficiently covered by the benchmarking exercise ⁽¹⁰⁾. However, the ranking has been useful as a political exercise, as it triggered a national debate on education and

⁽⁹⁾ The discussion on identification of relevant benchmarks is not new and not only linked to educational issues. The quote below points to reservations and illustrates how sensitive to the whole benchmarking exercise the identification of relevant features and benchmarks is. 'Companies often waste time benchmarking non-critical functions or struggling to raise the performance bar an insignificant amount. Does it really matter to your customers if you shrink your delivery time by two days? If you deliver packages, yes. If you deliver antiques, maybe not. And how much does it cost to cut out those days?' (Dahle, 1996).

⁽¹⁰⁾ 'In addition there are other important objectives. At primary and lower secondary school level one objective is to have pupils develop their creativity and self-confidence, as well as interpersonal and communication skills, and for the school to prepare pupils for life in a democratic society. These soft values need to be measured using new methods that are still in the developmental stage, though as yet there is still no sound body of data on which to make international comparisons. Indicators for the softer values are not included below, therefore'. *Strukturovervågning – International benchmarking af Danmark*, Finansministeriet May 1999, section 14.

helped identify areas for improvement or reorienting the focus of reforms. The benchmarking exercise was not used, therefore, to learn directly from others, but to stimulate a national debate by letting the figures speak for themselves.

The example points to the complexity of applying benchmarking to education systems. The approach presupposes that it would be possible to disaggregate a VET system into smaller units and that a benchmark could be singled out that could then be adopted by the institutional frameworks of other VET systems. However, as we have seen before, the highly context-linked nature of VET systems presents a major barrier for the benchmarking approach. The question is whether the benchmarking approach can be applied to social systems at all.

In a study for the European Commission on the possibility of benchmarking for employment policies (Schütz et al., 1998), a team of German researchers makes a distinction between benchmarking as an analytical inventory and as a policy tool for improvement. The first mainly implies a choice of benchmark areas for analysis and policy recommendations; the second is of a higher order and involves the understanding of performance gaps and the implementation, monitoring and evaluation of interventions. Benchmarking in the second sense has to be based on benchmarking as an analytical inventory.

The authors argue that, in principle, benchmarking for (employment) policies is possible, if a number of conditions are met:

- (a) a thorough understanding of the benchmarking approach;
- (b) a wide-ranging database and multi-method approaches;
- (c) adequate resources (time, money, personnel);
- (d) commitment of involved key actors to organisational learning.

They conclude that benchmarking is easier for organisational entities (such as employment offices) than various (employment) policies, whose success depend in part on other policies and institutions.

Grubb and Ryan (1999) see international or system level benchmarking as valuable but weakly experimental and consistent with fairly

weak research methods. Their comment is: 'It is more difficult to evaluate systems with their multi-dimensional goals and multiple intertwined characteristics, than it is to evaluate specific public programmes with typically unidimensional goals and simple attributes. It is also the case that no social experiment involving apprenticeship has ever been conducted. Consequently, while quasi-experimental evidence remains valuable, weakly experimental methods (notably international comparisons) become more important for the evaluation of apprenticeship than for other VET categories.'

The European Commission and the Member States' attempts to develop benchmarking procedures, following the Amsterdam Treaty goal to harmonise VET policies, are still in a very early stage. In employment policy, where the process is more advanced, the European Commission has defined benchmarking as: 'a formalised process by which employment performances of different countries are compared with each other in order to highlight the best performing ones, to set global targets for progress in the employment situation over the medium/long term and to identify which policies have been most effective in raising the level of performance' (European Commission, 1997).

The process implies:

- (a) setting verifiable objectives and deadlines, both at European and national level;
- (b) agreeing on common performance and policy indicators, based upon a country-comparable statistical basis;
- (c) undertaking peer reviews of major national employment policies and programmes.

Experience with the benchmarking of VET systems at EU level so far shows that the main problem is not so much the general acceptance, among Member States, of the approach, but the development of appropriate indicators and securing that data are available to measure these. However, there are also words of warning raised against believing that strategies of the best performers can be easily copied under different contexts. We have been aware of this danger during our previous evaluation and policy advisory work.

7. Conclusions and needs for further research

7.1. Conclusions

There is evidence of a shift in international review practice in VET from project-based to systemic approaches. This shift is related to the growing awareness of the restricted impact of single projects and their sometimes adverse effects on the system.

The aim of our study was to reach a better understanding of the underlying concepts and methods used for VET policy evaluations by exploring the contribution of research, as well as a number of examples from transition countries.

One of the central conclusions is that there is no holy grail in terms of conceptualisation or methodology related to VET policy evaluations. The engineer's toolbox is of limited use. Likewise, management approaches based on refined analytical frameworks often tend to obscure rather than illuminate VET, including reliance on tools derived from systems analysis. It appears that the only remedy is the evaluator's broad understanding of the essential components of VET, of their relationships, of the fundamental logic between the system and its environment and of strategic levers for change. This understanding develops only through many years of apprenticeship and first-hand experience of VET policy evaluations.

Using the hermeneutic approach, a particular system can be understood by alternating between studying parts (system elements) and totalities (the total system), as well as creating continuous interactions, in our interpretation work, between our preconception (*Vor-Verstehen*) and the new understandings worked out through the process. However, understanding always takes place between two entities both placed in a historical and cultural context, and Gadamer (1972) insists that the interpreter cannot abstract himself from personal history and culture. What the evaluator must do is openly recognise and clearly articulate this tension.

According to Habermas (1972), human science cannot stop short at interpreting and illuminating human actions and behaviour but, ultimately,

aims to equip individuals with an understanding of how best to pursue their objectives in a rational way.

We explored the question of whether we can provide a conceptual framework for grasping not only the parts (building blocks) of VET, but also its historic roots and dynamic relationships and, thus, for managing the dialogue with key actors from inside and outside the system. A number of essential functions and processes of a successful VET system and architectural elements were described which are, initially, sufficiently simple to engage in the dialogue. However, for developing a better understanding of VET, the building-blocks approach has to be accompanied by a deeper insight into the systemic logic of the VET system.

A structural-historical approach (*Strukturell-genetischer Zugang*) could be used. Here, the analysis would start with an overview of existing practices, functions and structures, as they appear to the (experienced) evaluator. Essential structures are then traced back to their origin (genesis), and an understanding/explanation of the historical context is reached in which the specific phenomenon was born or established.

The phenomenological/hermeneutic method is primarily used to achieve an understanding rather than for explanation, which is to state causes behind the phenomenon observed. There is a continuous interaction between understanding and explanation; they feed each other. Understanding is the point of departure for VET policy evaluations, and explanations are needed when we fail to understand. Useful explanatory frameworks include:

- (a) economic or labour market or sociological laws and other forms of determination of the concrete conditions under which a phenomenon is active;
- (b) functionalist explanations referring to the totality in which the phenomenon to be explained is placed and a description of the precise function it serves;
- (c) historical-cultural explanations behind the occurrence of different institutions and practices in contemporary VET systems;

- (d) structural(ist) explanations of phenomena which can only be uncovered through an in depth analysis;
- (e) system-analytical explanations unfolding how changes in one component will have spill over effects on other system components or on the total system (including discussions on internal and external consistency).

We conclude that, besides developing further the building-blocks approach by including explanations, and in order to avoid hobbyhorse explanations, we have to cultivate, individually and in a community of VET evaluation practitioners, a new attitude towards argumentation, a more open discussion practice. This we term a discourse.

VET policy evaluation requires a systemic understanding of VET. But how can the VET system universe be delineated? What relationships exist between the units? Which matrix of dominance patterns the interplay of units? What are the boundaries of the system? What relationship (metabolism) exists between the system and its environment? And what is the prime mover or strategic lever for change?

VET as a subsystem can be seen as an open system in a steady state; it depends on self-regulating mechanisms to maintain its boundaries and its continued existence within these boundaries (meaning primarily boundaries which distinguish the VET subsystem from other subsystems with the same society). The steady state depends on a balance of inputs and outputs. The inputs are demands and support: support makes the system strong enough to process demands and to produce outputs in the form of qualifications. The VET system sits in an environment – the total social system – and there is a continuous feed-back into the VET subsystem, signalling if outputs produce good or adverse effects in the environment. A systemic approach to VET evaluations focuses on the analysis of relationships, of communication channels, of responsiveness and adaptability, based on the fundamental understanding that changes in one component lead to changes in other components and in the system as a whole. So, it is important to understand the specific systemic logics of different VET systems. Hence, the interest in internal and external consistency, the preoccupation with systemic regulatory frameworks, and the search for the prime mover or the strategic lever related to VET system change.

Internal consistency indicates that improving VET (including increasing their attractiveness for both students and enterprises) can only be done through a system approach, encompassing all the components of the education system. The policy of a high degree of decentralisation and privatisation, and the modularisation of curricula, were discussed as two examples bearing the risk of lacking internal consistency.

External consistency refers to the relationships between educational objectives and the economic and social context. There is a range of potential problems connected to that: the employment system is not always able to articulate its needs; there are tensions between the pedagogical logic of school systems and competence needs of the employment system. Furthermore, the drive to develop national qualification systems as a means to increase internal and individual learner consistency may conflict with deregulated labour markets and short-term recruitment and training practices in companies. Finally, it is realised that education cannot by itself solve employment problems.

To balance system-inherent tensions, there is a need for some form of regulatory framework that would allow flexibility and adaptability while satisfying the need for high quality and social equality. Such a regulatory framework needs to guide reform efforts.

A systemic approach to VET evaluation, as described above, is the way forward. The ultimate aim of all VET reform endeavours is to better balance tensions and ensure internal and external consistency of the system. But there is no uniform model that could be applied.

Apart from the systemic approach, the evaluator needs to develop an understanding of how systems change in the interactions between actors, that innovations need to be taken forward not only to implementation but also to institutionalisation, and of the possible strategic levers for change.

Our experience with VET reforms in transition countries so far, as well as definitions of the term systemic reform, demonstrate that two aspects are key to the success of any systemic reform effort in VET:

- (a) the professionalisation and further development, in a structured and continuous way, of teaching and other education staff;
- (b) new school-university and local partnerships.

Change is a complex, multi-purpose, multi-layered, multi-actor, multi-stage process whose outcomes are partly unpredictable. System levels affected by each change comprise the primary learning processes, the organisational level, the institutional frame, and the policy and legal frame. According to Grootings (2002b), reform is constrained by the institutional context of the country, and there is an institutional logic that is country-dependent. Reform can only be undertaken through the mutation of historic institutional structures.

Another aspect that the evaluator must understand concerns suitable ways to gain understanding and support from the actors necessary for implementation. Accomplishing change is about reversing deeply embedded policies and strongly held beliefs. It is about preventing resistance to change or attenuating its impact.

System change is likely to be achieved through small, incremental change in narrow and targeted areas and only where there is equilibrium between radical change and traditional forces. Change requires, in a new political culture, a clear sense of public purpose, new partnerships and new skills, and careful policy coordination, compensatory mechanisms and collaboration in adequate forums for consultation and joint decision. The prevailing political culture will also determine whether reforms are implemented through top-down or bottom-up approaches (or a mixture thereof). Our experience has shown that contrasting organisational forms are required, i.e. a high degree of bureaucracy for structural reform and rather non-bureaucratic reforms for curricular innovation and the learning process.

Owing to the multitude of factors impeding on a system that is in a steady state of development and the unpredictability in particular of human behaviour, planning system change must always remain imprecise and its outcomes imperfectly defined. As a consequence, it becomes essential consciously to build into the change process strategic phases where social actors come together and reflect and learn with the ultimate aim of verifying or rectifying hypotheses as input into the next stage of the reform. This explains the need for evaluations not as one-off events, but as a continuous cycle.

Such evaluations are formative in nature, as they have influence in forming or developing poli-

cies. They are constructivistic in that they deny that there is an objective knowledge about the world (the VET system) and in that they put the actors at the centre of the evaluation rather than its outcomes. Nyhan (1998) outlines the lack of a research tradition in VET (as opposed to general education) and highlights the fruitful implementation of action research in education and training, which engages the social actors involved in transformation. The action research strategy (here used in its empowerment evaluation version) offers participants a voice in shaping the evaluation agenda, it involves an active feedback policy, and it sets a clear focus on utilisation and implementation. Action research offers channels for continuous dialogue, with the relevant stakeholders, about the validity of implementation of the concepts being studied, throughout the project.

This means that the process of undertaking an evaluation – i.e. the way in which it is negotiated and managed and how it impacts on its subjects – is itself seen as a legitimate object of study. It also follows that there is a thin line between evaluation and consulting.

Evaluations call for substantial resources in terms of staff, time and money. It is important to remember that, with continually evolving reform efforts, little time spent in the country and few cases explored, evaluations represent only a snapshot and general conclusions must always be considered tentative.

Policy evaluations require different evaluation criteria from project evaluation. These depend on the scope and purpose of the evaluation and should also be subject to negotiation with the actors involved.

There are no absolute standards for a good VET system, which the evaluator could use. Concepts and practices of VET systems of west European countries all have their positive and negative aspects; they cannot be detached from their institutional contexts. VET systems of west European countries have come under reform pressures themselves, as many fundamental problems have not been tackled, and new challenges have emerged. The adopted strategies used by west European countries for VET system renewal or reform reflect the different educational histories of each country rather than the problems these strategies are designed to address.

In these circumstances, a universal manual is difficult to assemble. EU experience has shown that neither the common factors pressing on national policy formation for education and training, nor the priorities that most Member States share, are necessarily leading to a uniform pattern of convergence (even if some common trans-European trends have emerged). The non-existence of clear-cut policy priorities or indicators may create a dilemma for the objective evaluator. A particular donor may impose them, but more easily on the evaluator than the specific government concerned. A possible way around the dilemma of preconceptions or nationally biased understandings is to start an evaluation by presenting the building blocks and formulating ideal practices against which actual practices would then be described and evaluated.

Benchmarking becomes a meaningful activity (or tool), when we view it as a formal and deliberate process of comparison aimed at detecting weaknesses and creating ideas for improvement. Benchmarking is distinguished from purely analytical methods of comparison in that specific performance indicators must be developed and units must be found which can be considered as best performers concerning the chosen indicators. The comparison then aims to find options to make progress towards the position of the best performer. However, we have to take seriously the warning that strategies of the best performers cannot easily be copied under different contexts.

A team of German researchers (Schütz et al., 1998) makes a distinction between benchmarking as an analytical inventory and as a policy tool for improvement. The first mainly implies a choice of benchmark areas for analysis and policy recommendations; the second is of a higher order and involves the understanding of performance gaps and the implementation, monitoring and evaluation of interventions. Benchmarking in the second sense has to be based on benchmarking as an analytical inventory. The latter is especially relevant for our purposes.

Essential conditions for using policy benchmarking include a thorough understanding of the benchmarking approach; a wide-ranging database and multi-method approaches; adequate resources (time, money, personnel); and commitment of involved key actors to organ-

isational learning. They conclude that benchmarking is easier for organisational entities than for policies, whose success depends on other policies and institutions.

Member States have accepted the approach, but there are problems in developing appropriate indicators and secure that data are available to measure these. There are also words of warning against believing that strategies of the best performers can be easily copied under different contexts. We have been aware of this danger during our previous evaluation and policy advisory work.

7.2. Needs for further research

What is presented in this study are the first results of what has turned out to be rather ground-clearing work in VET policy evaluation. Thus, we see our study as a basis for discussion and further research work. Further analysis is needed of the levers and critical factors for achieving genuine change, not only on paper but in the interactions between actors, which implies taking innovations to institutionalisation.

There is a need to develop both formative evaluation procedures and instruments to assist specific systemic reform initiatives in evaluating and monitoring progress towards goal achievement, and to inform future planning. An evaluation model must take into account the creation of new organisational arrangements attempting to bring about simultaneous renewal in more than one institution. Developing guidelines for long-term documentation and evaluation, which define needed data sources and indicators of development and achievement, would serve the continuing planning and development of such ventures and assist in developing comparable data to enhance our understanding of the limits and potential of specific reform efforts, not least the cost-effectiveness of various strategies. Further research is also needed to examine the real costs of education (or VET) reforms.

Given the importance of leadership in the success of these endeavours, additional research is needed to understand how leadership skills develop, and the strategies for developing them. When a study finds leadership to be a critical variable, it presents a dilemma for policy-makers.

Can such leaders be selected and/or trained, or do we have to wait for them to emerge? More research on the specific qualities of educational entrepreneurship is needed, particularly with a focus on how individuals can be trained to lead collaborative efforts.

Finally, more research is required into how benchmarking could be used to improve systems, which indicators should be selected for benchmarking VET systems on an EU or international scale, and what can be considered as best performers concerning chosen indicators.

List of abbreviations

ETF	European Training Foundation
NCEQ	Centre for examination and qualifications
OECD	Organisation for Economic Cooperation and Development.
PAS	Pedagogical Academy in Sarajevo
TIHR	Tavistock Institute of Human Relations
VET	Vocational and educational training

Bibliography and references

- Bengtsson, J. Labour markets of the future: the challenge to education policy makers. *European Journal of Education*, 1993, Vol. 28, Issue 2.
- Bîrzea, C. The dilemmas of the reform of Romanian education: shock therapy, the infusion of innovation or cultural 'decommunisation'? *Higher Education in Europe*, 1997, Vol. 22, No 3, p. 321-328.
- Boglia, L.; Parkes. *Report on VET in Bosnia-Herzegovina*. The World Bank/ETF – European Training Foundation joint project, 1996 [Unpublished draft].
- Bryant, I. Action research and reflective practice. In: Scott, D.; Usher, R. (eds) *Understanding educational research*. London: Routledge, 1996, p. 106-119.
- Carr, W.; Kemmis, S. *Becoming critical: educational knowledge and action research*. London: Falmer Press, 1986.
- Challis, R. E.; Mason, C.; Parkes, D. YTS and the local authority. *MSC Research and Development*, Vol. 37, 1985.
- Cort, P. *Vocational education and training in Denmark: brief description*. Luxembourg: Office for Official Publications of the European Communities, 2002 (Cedefop Panorama, series, 49).
- Dahle, C. Five uneasy pieces – Part 2: Benchmarking. *CIO Magazine*, 1 June 1996.
- Danish Ministry of Finance. *Strukturovervågning – International benchmarking af Danmark [Structural monitoring – international benchmarking of Denmark]*. Copenhagen: Danish Ministry of Finance, May 1999.
- Davies, N. *Europe: a history*. Oxford: Oxford University Press, 1996.
- Descy, P.; Tessaring, M. (eds) *Training and learning for competence: second report on vocational training research in Europe: executive summary*. Luxembourg: Office for Official Publications of the European Communities, 2001. (Cedefop Reference Series, 11).
- Dietzen, A. Fostering innovation by European research in VET. In: Dietzen, A.; Kuhn, M. (eds) *Building a European cooperative research tradition in vocational education and training*. Berlin: BIBB – Bundesinstitut für Berufsbildung, 1998b.
- Dietzen, A; Sellin, B. The impact of survey and analysis projects - towards a European dimension of VET. In: Dietzen, A.; Kuhn, M. (eds) *Building a European cooperative research tradition in vocational education and training*. Berlin: BIBB – Bundesinstitut für Berufsbildung, 1998a.
- Drucker, P. F. *Management challenges for the 21st century*. New York: HarperCollins Publishers, 1999.
- Durand-Drouhin, M; Bertrand, O. *Improving the responsiveness of vocational education and training: flexibility and/or consistency?* Paris: OECD, 1994 [Unpublished draft].
- Dyankov, A. *Current issues and trends in technical and vocational education*. Paris: Unesco, 1996. (Unevoc studies in technical and vocational education series, No 8).
- Elsenhans, H.; Nolke, A. *International networks: the case of the coordination of economic assistance to the CEEC and the NIS*. Leipzig: Institut fuer Politikwissenschaft, 1996 (Working paper).
- ETF. *A cross-country analysis of curricular reform in VET in central and eastern Europe*. Turin: ETF – European Training Foundation, 1999.
- ETF. *Baltic conference on VET: the importance of European cooperation*, Tallinn, 22-23 April 1996. Turin: ETF – European Training Foundation, 1996a.
- ETF. *Current constraints and challenges of the VET systems in the Baltic States: the importance of European cooperation*. Turin: ETF – European Training Foundation, 1996b.
- ETF. *Transnational analysis of VET in the new independent States and Mongolia*. Turin: ETF – European Training Foundation, 2000.
- ETF. VET, a donor coordination framework for donor coordination. Turin: ETF – European Training Foundation, 1997.
- European Commission / SCR (The joint service for the management of Community aid to non-member countries). *Guidelines for: drawing up terms of reference for evaluations, evaluation methodology, criteria and suggested layout for evaluation*

- reports. Brussels: European Commission, 1999 (SCR/F/5D(98)). Available from Internet: http://www.advd.unibe.ch/uls/evaluation/dokumente/EU_Evaluation_Guidelines_021999.pdf [cited 9.6.2003].
- European Commission, Directorate-General for Education and Culture. *European report on the quality of school education – Sixteen quality indicators*. Luxembourg: Office for Official Publications of the European Communities, 2000.
- European Commission, Directorate-General for Education and Culture. *A memorandum on life-long learning*. Brussels: European Commission, 2000 (SEC (2000) 1832). Available from Internet: <http://europa.eu.int/comm/education/policies/lll/life/memoen.pdf> [cited 9.6.2003].
- European Commission, Directorate-General for Employment and Social Affairs. *Commission draft for the joint employment report*. Brussels: European Commission, 1997.
- Fretwell, D. *A framework for evaluating VET*. Turin: ETF – European Training Foundation, 2000 [Unpublished draft].
- Gadamer, H. G. *Truth and method*. London: Sheed and Ward, 1981.
- Gadamer, H. G. *Wahrheit und Methode: Grundzüge einer philosophischen Hermeneutik* (3rd ed.). Tübingen: JCB Mohr (Paul Siebeck), 1972.
- Gasskov, V. *Managing vocational training systems: a handbook for senior administrators*. Geneva: ILO, 2000.
- Grootings, P. *A comparative review of VET and labour market developments in south eastern Europe*. Turin: ETF – European Training Foundation, 2001 (Working paper).
- Grootings, P. *A note on evaluation*. Turin: ETF – European Training Foundation, 2000 [Unpublished draft].
- Grootings, P. *Advisory forum 2003 discussion paper*. Turin: ETF – European Training Foundation, 2002a (Internal paper).
- Grootings, P. *VET in transition: an overview of changes in three east European countries*. *European Journal of Education*, 1993, Vol. 28, No 2.
- Grootings, P. *VET reform as institutional change: looking at VET reform through the eyes of New Institutional Economics*. Turin: ETF – European Training Foundation, 2002b [Unpublished draft].
- Group of 40 Key Actors: Green Paper: *Strategy and policy for reform of VET in Bosnia-Herzegovina*, EU Phare VET programme, published locally, 2000.
- Grubb, N.; Ryan, P. *The roles of evaluation for VET: plain talk on the field of dreams*. Geneva: ILO – International Labour Organisation, 1999.
- Habermas, J. *Knowledge and human interest*. London: Heinemann, 1972.
- Hindle, T. What is a strategy? In: *Economist Books – Pocket strategy*. London: Penguin, 1994.
- Hodgson, G. Socio-economic consequences of the advance of complexity and knowledge. In: *Future trends 7: an information base for scanning the future* (CD-Rom). Paris: OECD, 2001/02.
- Howse, G. VET under review: the challenges in central Asia. *European journal of education*, Vol. 36, No 1, 2001.
- Jaworski, B.; Phillips, D. (eds) *Comparing standards internationally: research and practice in mathematics and beyond*. Oxford: Triangle Journals, 1999, Vol. 9, No 1 (Oxford studies in comparative education series).
- Kenneke, L. *A guide for evaluation of technical and vocational education curricula*. Bonn: Unesco/Unevoc, 1987 (Unevoc studies in technical and vocational education series, No 3).
- King, K.; Buchert, L. *Changing international aid to education: global patterns and national contexts*. Paris: Unesco /Norrag, 1999.
- Köhler, W. Gestalt Psychology Today. *American Psychologist*, 1959, Vol. 14, p. 727-734.
- Kuhn, T. S. *The structure of scientific revolutions*. Chicago: University of Chicago Press, 1996 (3rd ed.).
- Lasonen, J. (ed.) *Reforming upper secondary education in Europe*. Jyväskylä: Institute for Educational Research, 1996.
- Lasonen, J.; Manning, S. How to improve the standing of vocational compared to general education: a collaborative investigation of strategies and qualifications across Europe. In: Descy, P.; Tessaring, M. (eds) *Training in Europe: second report on vocational training research in Europe 2000: background report*. Luxembourg: Office for Official Publications of the European Communities, 2001, Vol. 1, Part 1, p. 115-167. (Cedefop Reference series, 3008).
- Lasonen, J.; Young, M. (eds) *Strategies for achieving parity of esteem in European upper secondary education*. Jyväskylä: Institute for Educational Research, 1998.

- Luburic, V. et al. *Teachers' and Trainers' Training in VET*. Zagreb: Croatian National Observatory, 2001.
- Luhmann, N. *Soziale Systeme. Grundriss einer allgemeinen Theorie*. Frankfurt: Suhrkamp, 2001.
- McDonald, R. et al. *New perspectives on assessment*. Paris: Unesco 1995 (Unevoc studies in technical and vocational education series, No 4).
- McLeish, E. A.; Phillips, D. (eds) *Processes of transition in education systems*. Oxford: Triangle Journals, 1998, Vol. 8, No 2 (Oxford studies in comparative education series).
- Mebratu, T.; Crossley, M.; Johnson, D. (eds) *Globalisation, educational transformation and societies in transition*. Oxford: Symposium Books, 2000.
- Miegel, M.; Nölke, A. *Transformation strategies and the role of economic assistance to central and eastern Europe*. Leipzig: Institute for Political Science, 1996.
- Montenegrin National Observatory. *Report on the VET system – Montenegro*, Turin: ETF – European Training Foundation, 2001.
- Nielsen, S. *Report on the evaluation of the EU Phare 1995 VET reform programme in Romania*. Turin: ETF – European Training Foundation, 1999a.
- Nielsen, S. *Report on the evaluation of the EU Phare 1995 VET reform programme in Estonia*. Turin: ETF – European Training Foundation, 1999b [Unpublished draft].
- Nielsen, S. *Report on the evaluation of the EU Phare 1998 VET programme in Bosnia-Herzegovina*. Turin: ETF – European Training Foundation, 2001.
- Nielsen, S. *VET teachers' and trainers' training in Croatia: challenges, priorities and short-term assignments*. Paper submitted to the workshop on 'VET policies' held in Split, Croatia, on 30 January 2002. Turin: ETF – European Training Foundation, 2002 [Unpublished draft].
- Nölke, A. Implementing economic assistance for central and east Europe: the problem of coordination. In: Nölke, A.; Miegel, M. (ed.) *Transformation strategies and the role of economic assistance to central and eastern Europe*. Leipzig: Leipziger Universitätsverlag, 1996, p. 171-192.
- Nölke, A.; Miegel, M. (ed.) *Transformation strategies and the role of economic assistance to central and eastern Europe*. Leipzig: Leipziger Universitätsverlag, 1996.
- Nyhan, B. Promoting a European VET research tradition: the role of the survey and analysis measures of the Leonardo da Vinci Programme. In: Dietzen, A.; Kuhn, M. (eds) *Building a European co-operative research tradition in vocational education and training: the contribution of the Leonardo da Vinci programme's surveys and analyses*. Berlin: BIBB – Bundesinstitut für Berufsbildung, 1998.
- OECD. *Development cooperation: efforts and policies of the members of the development assistance committee*. Paris: OECD, 1997.
- OECD. *Education and skills catalogue*. Paris: OECD, 2001.
- OECD. *Education policy analysis*. Paris: OECD, 1999-2001.
- OECD. *Employment outlook*. Paris: OECD, 2001.
- OECD. *Evaluating and reforming education systems*. Paris: OECD, 1996.
- OECD. *From initial education to working life: making transition work*. Paris: OECD, 1996.
- OECD. *Future trends 7: an information base for scanning the future* (CD-Rom). Paris: OECD, 2001/02.
- OECD. *Implementing the OECD job's strategy: lessons from member countries' experience*. Paris: OECD, 1998.
- OECD. *Motivating students for lifelong learning*. Paris: OECD, 2000c.
- OECD. *Reviews of national policies for education: Romania*. Paris: OECD, 2000b.
- OECD. *The appraisal of investments in educational facilities*. Paris: OECD, 2000a.
- Paquet, G. The new governance, subsidiarity and the strategic state. In: *Future trends 7: an information base for scanning the future* (CD-Rom). Paris: OECD, 2001/02.
- Parkes, D. (ed.) VET under review (editorial). *European Journal of Education*, 2001, Vol. 36, No 1.
- Parkes, D. Comparative intervention in the reform of VET. *European Journal of Education*, 1995a, Vol. 30, No 3.
- Parkes, D. et al. *A cross country analysis of curricular reform in vocational education and training in central and eastern Europe* / European Training Foundation. Luxembourg: Office for Official Publications of the European Communities, 1999.

- Parkes, D. Management of change. *European Journal of Education*, 1991, Vol. 26, No 1.
- Parkes, D. Management of change. *European Journal of Education*, 1997, Vol. 32, No 3.
- Parkes, D. Parity of esteem. *European Journal of Education*, 1993, Vol. 28, No 2.
- Parkes, D. Recent trends in vocational education and training. *European Journal of Education*, 1995b, Vol. 30, No 3.
- Parkes, D. The international perspective. In: Stanton, G.; Richardson, W. (eds) *Qualifications for the future: a study of tripartite and other divisions in post-16 education and training*. London: FEDA – Further Education Development Agency, 1998, p. 171-188 (FEDA report series, Vol. 2, No 5).
- Parkes, D. VET in transition, renewal, reform and reconstruction. *European Journal of Education*, 1999, Vol. 34, No 2.
- Ryan, P. *The school-to-work transition: problems and indicators*. Cambridge: Cambridge University Press, 2001.
- Schütz, H.; Speckesser, S.; Schmid, G. *Benchmarking labour market performance and labour market policies: theoretical foundations and applications*. Berlin: WZB – Wissenschaftszentrum für Sozialforschung, 1998 (Discussion Paper FS198-205).
- Sellin, B. et al. *Scenarios and strategies for vocational education and training in Europe: European synthesis report on phase I*. Amsterdam: Max Goote Kenniscentrum, 2000.
- Shaw, G. European standards in VET: what are they and who wants them? *European Journal of Education*, 1999, Vol. 34, No 2.
- Stanton, G.; Bailey, B. VET under review in England: trends and developments. *European Journal of Education*, Vol. 36, No 1, 2001.
- Stern, E. *Development evaluation: learning from social and organisational innovations*. London: TIHR – Tavistock Institute of Human Relations, 1989.
- Stern, E. *Philosophies and types of evaluation research*. In: Descy, P.; Tessaring, M. (eds) *Third research report: background report*. Luxembourg: Office for Official Publications of the European Communities, 2003. (Cedefop Reference series - forthcoming).
- Stratmann, G. *Help me! (But how?). Economic assistance to central eastern Europe in relation to transformation theory*. Leipzig: Institut für Politikwissenschaft, 1995 (Working paper).
- Stratmann, G. *Sectoral coordination of international assistance to central and eastern Europe 1990/1996: the case of environment and education*. Leipzig: Institut fuer Politikwissenschaft, 1997 (Working paper).
- Stratmann, G.; Nolke, A. *The design and implementation of economic assistance to central and eastern Europe: suffering from a lack of integration and coordination*. Leipzig: Institut fuer Politikwissenschaft, 1996 (Working paper).
- Teichler, U.; Gordon, J.; Maiworm, F. (eds) *Socrates 2000 evaluation study*. Paris: EIESP – European Institute of Education and Social Policy.
- Tessaring, M. *Training for a changing society: a report on current vocational education and training research in Europe* (2nd ed.) Luxembourg: Office for Official Publications of the European Communities, 1999. (Cedefop Reference Series).
- The World Bank. *Hidden challenges to education systems in transition economies*. Washington, DC: the World Bank, 2000.
- The World Bank. *Labor markets and social policy in central and eastern Europe: the transition and beyond*. Oxford: Oxford University Press, 1994.
- UNDP. *Human development report: Bosnia and Herzegovina*. New York: United Nations Development Programme, 1998.
- Unesco. *Technical and vocational education and training: a vision for the twenty-first century*. Set of recommendations to the Director General of Unesco at the Second International Congress on Technical and Vocational Education, Seoul, April 1999.
- Unesco/Unevoc. *Cooperation with the world of work in technical and vocational education*. Berlin: Unesco/Unevoc, 1998.
- Unesco/Unevoc. *Vocational education and training in Europe on the threshold of the 21st century*. Berlin: Unesco, 1999 (Unevoc document No ED/IUG/018).
- US Department of Education. Background of the study: study aims and research questions. In: *Systemic reform in the professionalism of educators*. Washington, DC: US Department of Education, 1995. Available from Internet: <http://www.ed.gov/pubs/SER/ProfEd/bkgd.html> [cited 9.6.2003].

- Van Wieringen, F. Scenario planning for vocational and adult education. *European Journal of Education*, Vol. 34, Issue 2, 1999.
- Viertel, E. Input into the report on the Peer review of initial VET in Croatia. Turin: ETF – European Training Foundation, 2003 (*forthcoming*).
- Viertel, E.; Grootings, P. *Review and lessons learned of Phare vocational education and training reform programmes 1993-1998*. Turin: ETF – European Training Foundation, 2000.
- von Bertalanffy, L. An outline of general systems theory. *British Journal for the Philosophy of Science*, 1950, Vol. 1, No 2.
- Watson, K. (ed.) *Doing comparative education research: issues and problems*. Oxford: Symposium Books, 2001.
- Whitman, I. *The OECD approach to education policy reviews*. Paper delivered to the workshop on 'Methods and tools for the assessment of VET systems', ETF – European Training Foundation, Turin, 26-27 November 2002. Paris: OECD, 2001 [Unpublished draft].
- Widmaier, A. *L'évaluation de L'investissement en formation professionnelle*. Turin: ETF – European Training Foundation, 2000.
- Willie, E.; Hodgson, P. *Making change work*. Cirencester, UK: Management Books 2000, 2001.
- Windham, D. M. *The OECD thematic reviews of education policy in south eastern Europe: a proposed outline for authors of country chapters*. Paris: OECD, 2000 [Unpublished draft].
- Wolf, A. Some final thoughts: vocational education policy in a European context. *European Journal of Education*, 1993, No 2.
- Wurzburg, G. *A systemic approach to evaluating the implementation of lifelong learning in OECD member countries*. Paper delivered to the workshop on 'Methods and tools for the assessment of VET systems', ETF – European Training Foundation, Turin, 26-27 November 2002. Paris: OECD, 2001 [Unpublished draft].
- Young, M. *Contrasting approaches to the role of qualifications in the promotion of lifelong learning*. Paper delivered to the workshop on 'Methods and tools for the assessment of VET systems'. ETF – European Training Foundation, Turin, 26-27 November 2002. London: Institute of Education, 2001 [Unpublished draft].
- Young, M. Improving vocational education: trans-European comparisons of developments in the late 1990s. In: Stenström, M.-L.; Lasonen, J. (eds) *Strategies for reforming initial vocational education and training in Europe*. Jyväskylä: Institute for Educational Research, 2000.



Look, listen and learn: an international evaluation of adult learning

Beatriz Pont, Patrick Werquin ⁽¹⁾⁽²⁾

Abstract

This article analyses the process involved in, and the impact of, an international evaluation of adult learning policies and practices conducted by the OECD's Education and Training Policy Division: this thematic review includes both national and international quantitative and qualitative analysis, and field visits to nine OECD countries. The article gives an overview of the review process as a whole, including its main results, and examines some of its impacts and outcomes. The analysis undertaken, the interest generated in the countries, the consistency between recommendations made and recent reforms in most participating countries, and the undertaking of a second round of visits in some countries, show that the objectives have been achieved. The thematic review seems to have had an impact on adult learning policy agendas and is now a benchmark for research and for better knowledge of good practices in this field.

⁽¹⁾ The authors would like to thank Sylvie Chioussé and Pascaline Descy for their comments on the previous versions.

⁽²⁾ This paper is the responsibility of the authors and does not necessarily reflect the views of the OECD or the governments of its Member Countries.



Table of contents

1. Introduction	240
2. Reasons for a thematic review of adult learning	242
2.1. Supplementing the quantitative approach: how field visits contribute	242
2.2. Adult learning as a subject	243
2.3. Observing all the parameters of adult learning	244
3. Process and method: interacting with countries and meeting their expectations	245
3.1. The background report: describing	245
3.2. The visits: seeing and understanding	246
3.3. The country note and the comparative report: analysing	248
4. A range of evaluation results	250
4.1. Positive country responses	250
4.2. Raising awareness: adult learning is a key policy issue	252
4.3. High quality evaluation by outside experts may trigger action	252
4.4. Improved national and international cooperation	253
4.5. Some general recommendations from the thematic review	254
4.6. Recommendations specific to countries reveal similar problems and issues	255
5. Comments and prospects – an exercise worth pursuing	258
List of abbreviations	260
References	261

List of tables

Tables

Table 1: Recent reforms of adult learning policy	250
Table 2: Country recommendations by the OECD teams of experts	256

1. Introduction

Adult learning ⁽³⁾ is finally a key concern of policy-makers and researchers for a broad range of reasons ⁽⁴⁾. These include population ageing, the scarcity of some skills and/or qualifications, the upsurge in new information and communication technologies in all areas of work and life, and the need to provide OECD populations with a mastery of basic skills (reading, writing, numeracy) that is sufficient for them to exercise their rights and duties as citizens to the full.

In this field, it is customary to use two methods of evaluation. One is quantitative, using surveys or the like: data are collected in order to pinpoint any systematic trends and/or characteristics of the past that may offer – or so it is hoped – clues to the future. The other is more qualitative: data collected on a large scale are verified (compared) using more qualitative information so that interpretation of the figures does not lead to explanations that are too remote from the actual situation. Qualitative surveys also make it possible to gain a better understanding of the players and, through interviews of all kinds and formats, to obtain information that can be used to evaluate the actual situation.

In terms of quantitative evaluation of adult education and training, the international adult literacy survey (IALS) made a clear contribution as a result of the shock waves that it generated in the countries taking part: even among the leading countries, a quarter of the adult population did not possess the minimum skills needed to function fully in their country of residence (OECD and Statistics Canada, 2000). In some ways, this survey marks the start of a major new awareness. In countries such as the UK (Moser, 1999) or in the Scandinavian countries – given as an example because many reforms are underway there – programmes and laws geared towards improving the population's basic skills have proliferated since then.

In the case of qualitative evaluation, the aim is to measure the impact of programmes that are under way, to locate good practices and suggest improvements. This is the role of the OECD's thematic review of adult learning. It combines qualitative evaluation with field visits to the participating countries to provide an integral and broad evaluation of policies, practices and results.

This article looks at the information that can be used to evaluate the evaluation method represented by a thematic review. More precisely, it attempts to pinpoint whether, and the ways in which, a thematic review, as conducted at present by the OECD, can have an impact on measures taken in the review countries. More generally, the article, as it is designed, is intended in particular to find out whether the thematic review approach generates new impetus likely to bring about reforms in the country visited. In this respect, one consideration to be made is that the OECD's thematic review is voluntary in nature: each country is free to decide whether or not it wishes to take part. Evaluation of the correct operation of reforms, therefore, has to take account of a mixture of causes and effects:

- (a) the country has already launched reforms and the thematic review merely observes and approves or does not approve the reforms; the impact of the thematic review is probably marginal in such cases, but it may help to evaluate or validate the reforms under way;
- (b) the country is in a period of preparation and analysis, with the result that the thematic review recommendations are likely to have more of an impact;
- (c) or, most often the case, the two are mixed: the task of the thematic review experts is both to judge any reforms under way and to think about what could be made to work better. Pinpointing reasons for success or

⁽³⁾ The term 'adult learning' should be seen as a synonym of 'adult education and training'. The concept encompasses different reasons for learning (personal, professional) and the methods used (formal, informal or non-formal). It covers all the activities through which adults learn after they have left initial education and training.

⁽⁴⁾ Previous works have paid much more attention to other areas of education and training; primary, secondary and higher education, transition from school to work and even, more recently, preschool education.

failure, or assessing the validity of the recommendations made, then becomes problematic as a whole range of causes and effects are mixed up at the time of the evaluation.

One of the comments made most often during the field visits was that the visit by the OECD experts came at a bad time and, at the same time, at a good time. It came at a bad time because the country was in the throes of reforming its adult learning system and at a good time for exactly the same reason. The thematic review evaluation approach, therefore, was also used for immediate evaluation purposes since it provides a running commentary on the current situation by outside visitors who hold a mirror in which the country's players can see themselves and pinpoint good and bad choices. It needs, nevertheless, to be supplemented by a second visit, once the reforms have been completed, to make up for the 'you have come too early' aspect.

There are some remarks to be made with regards to the timing of the reviews. First of all, there is the time needed for reforms and the time needed for evaluation: they are different. The first process is fairly long in itself (four years to prepare the law on continuing vocational training in Switzerland, for instance), but the time needed for evaluation is probably even longer and it might seem somewhat daring to start talking about evaluation before reforms have filtered down to field level and to practitioners (in the UK, it is hoped to halve the number of people in the

two lowest literacy levels by 2015). Nevertheless, the speed at which reforms are disseminated and their conscientious application are interesting factors to study in a context of global evaluation.

All the countries, without exception, invited the OECD to review their education and training system because they were preparing laws or reforms in this area. The thematic review, therefore, was very well-timed but was not the catalyst for a major policy of reform. It did not have the potential of the IALS, which came at a time when problems of low literacy levels were not really at the policy forefront, in contrast, for instance, to young people's unemployment.

Finally, the countries visited tend to work with a fixed education and training budget, whatever the age and level of the target group (from preschool to university of the third age). The interest, therefore, is to find out whether a balancing act is involved – what is given to one sector is taken from the budget of another, as was the case in Finland with the greater investment in preschool education – or whether supplementary funds are released. Any such discussion might nevertheless remain in the realms of theory as it is very difficult to obtain precise enough data on expenditure broken down by item.

This article looks at the process, findings and impact of the thematic review on adult learning in the countries which took part in the first stage of the review: Canada, Denmark, Finland, Norway, Portugal, Spain, Sweden, Switzerland and the UK.

2. Reasons for a thematic review of adult learning

There has been fresh interest in recent years in adult learning – which has references to it in very old texts such as those of Plato or Condorcet – because there is now little doubt that human capital has an impact on economic growth. People's education and training are, moreover, humanist concerns which have always existed, whether or not they have been placed on an institutional footing, irrespective of the economic dimension. Adult populations, almost by definition, are outside the framework of initial education and training: levers have to be found to reach them and to implement measures to provide basic skills or prevent skills and knowledge from becoming obsolete. This is as true in the context of the industrial restructuring which is currently taking place as it is in the context of a personal wish for advancement and/or new directions. The clear message from international organisations (Unesco, OECD, European Union), and taken up by the national authorities, is that lifelong learning – and therefore, as a subset of this broad concept, adult learning – is central to current concerns. Lastly, the growing importance of the information society and economy in professional work and in daily life means that many people need to learn in order to keep abreast of the new information and communication technologies.

2.1. Supplementing the quantitative approach: how field visits contribute

There are various ways in which a set of questions, such as those surrounding adult learning, can be analysed. One is quantitative: it is always

possible to capture, in an adequate database, factors relevant to what one is trying to improve or explain. The incidence of participation in adult learning is the most obvious. If these factors can be measured ⁽⁵⁾, it is possible in practice to try to explain their variations over time for the same person ⁽⁶⁾ or between people at the same time, or both ⁽⁷⁾. They are then variables explained by, or typically independent from, a statistical or econometric model. It is also possible, although often much more complicated, to pinpoint the explanatory variables causing variations in the variable to be explained ⁽⁸⁾. They form the counterpart to the statistical model, although if this model is to be well constructed, many of these explanatory variables will have to be monitored or techniques used to make up for their absence (monitoring for non-comparability which is not observed). A quantitative analysis should therefore make it possible to pinpoint those factors, causing variations in the variable of interest: in this case, participation by adults in learning activities. Although such approaches are clearly of interest (they are relatively simple to implement, do not cost much if the data are available and the same analyses can be infinitely reproduced in different contexts or for different sub-populations) their drawbacks are just as clear. One prerequisite is obviously that the data have to have been collected, whether intentionally or not. If these data have not been intentionally collected, other conditions have to be satisfied: monitoring of appropriate variables, correct measurement of important variables, to cite only the most important, without forgetting the well-known problems of sample size, representativeness, missing data, etc.

However satisfactory they may be, quantitative methods cannot, therefore, be the only founda-

⁽⁵⁾ This database may, for instance, contain a variable representing the number of hours spent in training by the persons surveyed, or any other measurement of the intensity of commitment to training activities.

⁽⁶⁾ The unit of observation may be, fairly neutrally, a region or a country.

⁽⁷⁾ This requires data still rather rare in education – even though increasingly common in labour economics or health, for instance constructed from questioning of the same people at regular time intervals (a year for example) over a relatively long period (10 years seems to be the minimum) to ensure that any variations observed are significant.

⁽⁸⁾ Since they cannot all be observed as their measurement is tricky or impossible.

tion for decision-making ⁽⁹⁾. The nature of some magnitudes – financing, for instance: real cost, expenditure broken down accurately for the population's different target groups – is very hazy for education in general, and adult learning in particular. It is clear, therefore, that few decisions are genuinely based on adequate statistical modelling which provides a convincing picture of the actual situation. The observation of figures has to be supplemented – and possibly even substituted, if there is no convincing database – by more qualitative work.

The second method, therefore, is qualitative and this work may itself take a variety of forms. One of the best documented involves detailed surveying – by interviews which may or may not be direct – of a small number of people. These are, however, techniques which are more geared to individual problems, for instance at the level of an enterprise or a particular geographical area. It is less easy to use them for a country, even less so for international comparison.

Adult education and training, in the context of lifelong learning, is a fairly new field which is not very cohesive. It is fragmented, decentralised, and has a large number of service providers. An evaluation based on quantitative data and bibliographical research is not enough. It is necessary to visit the field, to visit the various providers and to gain a picture of all the issues involved in policy-making. It is only in this way that it can be hoped to understand discussions on adult learning. This is why this approach is useful for an accurate evaluation of adult education and training programmes.

The OECD Secretariat, and its Education and Training Policy Division in particular, has extensive experience of this type of work which is referred to as a thematic review. It may take other forms and names – peer review, for instance – and the following analysis undoubtedly applies to all studies which closely combine, by construction, theoretical thinking, quantitative data and field visits (for the qualitative dimension).

In the case of a thematic review of adult learning, the main aim of the exercise is to find out under what conditions different policies or

approaches may lead to better outcomes, and to provide lessons for the future and for the countries being compared. In other words, it tries to identify good practices and to measure their transposability. A complex, multi-stage, process has been developed for this purpose. The purpose of each stage is to highlight what has led an experiment to bear fruit; either it has made access to training easier for adults or has directly helped to increase participation by adults in learning activities.

2.2. Adult learning as a subject

One of the most obvious findings of the thematic review is that there is no systematic evaluation of adult learning systems or programmes in the review countries. Even in countries in which the adult learning system is of high quality and good results are being achieved, it seems that evaluation is not a high priority. There are many reasons for this, the most important of which is undoubtedly the very complex nature of the adult learning universe:

- (a) many institutions and actors are involved in this system;
- (b) there is a striking lack of coordination;
- (c) it is difficult to implement a convincing evaluation strategy;
- (d) adult learning has an impact in many fields (some of these external effects – in the private sphere for instance – are difficult to pinpoint and little is known about them) and very often in the long term.

It is also for all these reasons that adult learning is a good area for experimentation to find out what contribution field surveys can make to knowledge of issues and achievements, enabling policies to be developed in a more coherent and more efficient way.

Furthermore, there are still major gaps in participation in adult learning and another reason for this thematic review is to identify these deficits and the best ways in which they can be resolved in the particular context. The starting

⁽⁹⁾ Except perhaps in relatively elementary cases in which decisions are based less on statistics than on enumeration work: a census gives a fairly direct picture of the number of schools that need to be built or the number of teachers that need to be recruited.

point of this review is that a comparative international analysis of national adult learning policies and practices, based on field visits, is a valuable tool for identifying good and less good practices – even if they are not immediately transposable – as well as the lessons learned from attempts to modernise and adapt systems.

2.3. Observing all the parameters of adult learning

A thematic review as conducted by the OECD is considered by its member countries to be an appropriate tool for understanding the issues involved in the field in question as it brings together all the dimensions that need to be considered. Without dwelling on detail, a thematic review makes it possible to:

- (a) identify and take stock of the political, social, economic and cultural context in which adult learning is taking place – explanations for good and bad practices are largely to be found here;
- (b) visit a wide range of institutions and actors of adult learning – government agencies, employers, trade unions, teachers, learners – and to measure their motivations for involvement: it is undoubtedly these actors who are looking for changes and new measures;
- (c) identify the objectives set for adult learning: improving human capital, new and/or specific qualification needs, increase in the active population, creation of partnerships – these objectives shape any incentives that need to be found, whatever at macro (country), meso (institutions) or micro (individuals) level;
- (d) gauge political commitment to the adult learning process and locate attempts to formalise demand from the bottom up and decentralised initiatives;
- (e) measure the impact of schemes and programmes for end users: learners, employers, various members of the community, different organisations and the country as a whole;
- (f) find out whether there is a culture of evaluation of adult learning schemes and, if so, to find out what methods are being used. Are they comprehensive? Do they produce relatively objective indicators?

- (g) find out if any *ex ante* or concomitant evaluations are having an impact on adult learning programmes.

The unit of time is also a key dimension. It enables international comparison and a better evaluation of findings and progress. Although it is not at all necessary to set an identical time frame for all countries – as the magnitudes being studied, such as the basic competences of adults, do not evolve very quickly – it is certainly necessary for the study to be conducted within a reasonable time frame (two or three years after the reforms). Secondly, the fact that the countries reviewed share the same enthusiasm, or the same upturn in interest, at the same time, can facilitate evaluation and comparison, especially in light of the role that convinced actors can play in the success of the activity.

Another important dimension is the geopolitical context common to some groups of countries which makes it possible to compare them with, as a criterion, a similar context. This is particularly true of the Member States of the EU (six of which are included in the first round ⁽¹⁰⁾) or, to a lesser extent, the member countries of the OECD; in contrast with Unesco, for instance, which is a much larger and clearly less comparable grouping. In both the EU and the OECD the rules of economic operation are fairly comparable and there is a certain political stability. However, and it is here that the interest of the study lies, there is no guarantee that adult learning cultures and practices are comparable. It is up to the experts conducting the thematic review to highlight relative similarities or differences as well as the consequences, whether good or bad, that they may have.

The only dimension that is genuinely missing in this picture is the link that might exist between the dissemination of the thematic review's results and findings and the schemes and programmes set in motion subsequently in the country. It could then be interesting to pinpoint the level at which such schemes are run: local or national, piecemeal or global? This picture can be supplemented by other information – the extent to which findings are disseminated to the actors, speed of reaction in terms of the launch of new schemes, for instance – but this is not really feasible unless there is a second visit to the countries after a few years.

⁽¹⁰⁾ Eight are members of the European Free Trade Association (EFTA).

3. Process and method: interacting with countries and meeting their expectations

The review process is a mixture of descriptive work by the countries, and analysis by the experts making the field visits. A document (background report, country note and comparative report) is drawn up for each main stage. This is justified by the complex nature of the problem and the need for interaction with the actors of the countries visited.

3.1. The background report: describing

The purpose of the background report is to provide the experts responsible for reviewing the countries with the information that they need to be able to carry out the review. It may take the form of a brief history of adult learning ⁽¹¹⁾, a description of programmes under way and policies being implemented or planned, etc. The experts selected to make the visit and draft the country note, although specialists in the field, do not necessarily know the country to be reviewed or about recent measures in the adult learning field.

The background report is not comparative, and is not intended to be so, even though it is used as a basis for drafting the final comparative report (OECD, 2003). Its purpose is to report on the present adult learning situation and therefore to provide the following information:

- (a) the context of, and stage reached with, thinking about adult learning in the country;
- (b) a breakdown of access;
- (c) learning motivations;
- (d) measures to bring the training supply and demand into step;
- (e) adults' needs and the way in which they are being satisfied;

- (f) public action at all levels, whether national, local or municipal;
- (g) evaluation, whether economic or social, individual or collective, recent initiatives and their respective impact.

This background report is intended to be descriptive and is drawn up by the country being reviewed.

It has to be descriptive to ensure that the conclusions of the subsequent analysis retain a degree of neutrality. The information is supplied by the country being reviewed. A clinical report on what has happened, is happening and, in some cases, will happen ⁽¹²⁾ in the adult learning field means that the experts are not tempted to draw overly hasty conclusions. Analysis is their task and is contained in the country note.

The background report needs to be drafted by the country being reviewed for two main reasons. First, these two reports – the background report and the country note – make it possible to pinpoint any contradictions between the internal vision and what may be perceived from outside. In addition, and in particular, it helps more certainly to pinpoint the real variables explaining the process in question: i.e. access to and participation in adult learning. In practice, if the approach described above which underpins a thematic review is correct – i.e. finding out about, partly through field visits, the reasons that make adults look for and take part in learning activities – leaving the drafting of the background report to the countries visited helps to improve the chances of locating the real factors – deeply rooted in history, culture or the system – behind what is observed during the visit. The quantitative data contained in the annex to the Background Report or collected elsewhere (IALS or ELFS

⁽¹¹⁾ Without prejudging the conclusions, it is already clear that some countries have a long tradition in this field – Denmark, for instance, with the pioneering thinking of Grundtvig – and that more recent events, such as the Carnation Revolution in Portugal, have shaped the systems of other countries.

⁽¹²⁾ It is fascinating to note, in some cases, that the preparations for a law may be so charged that they influence current practices: this was clearly the case in Switzerland, where the visit took place several years before the law on continuing vocational training scheduled for 2003-2004.

(European labour force surveys)) can be fleshed out only by an approach of this type.

Countries receive common guidelines on the content of the background report, set out in the review's terms of reference. These guidelines are intended to help the country in its search for information and data on adult learning. They are also intended to ensure – even though it is not compulsory to answer all the questions and countries have considerable discretion in presenting their research as they wish – that the reports from the various countries have enough in common to facilitate comparison). It should also be stressed that depending on the country, the background report is commissioned from researchers to the academics or drafted internally in a ministry. In any case, work always entails cooperation between a number of people or organisations. It is very rare for the background report to be controversial in the country and this contributes to its value.

The relative consensus generated by the background report is often due to the fact that it is one of the few exhaustive documents drafted on adult learning. There are two main reasons for this. First, adult learning is a relatively new field and few institutions, apart from ministries, can undertake such work. The second reason has to do with the nature of this work. It is rare for a country to be able to lay claim to an integrated adult learning system, and it is more likely that there is a profusion of schemes, programmes and initiatives supplementing or competing with one another. Drafting the background report, therefore, is a very complicated exercise for which local researchers may lack motivation and which can take place only in partnership. However, an activity such as an OECD thematic review can generate enough incentive, or even enthusiasm, for the production of a document which, in some cases, is also independently published in the country involved. The background report is often the first attempt to survey all the schemes undertaken in the country at the initiative of various ministries, social partners, non-governmental organisations (NGOs) or others (associations, sponsors, etc.). It is also often the first time that a sum of knowledge and data is assembled in a single document in a format accessible to a wide range of readers.

The background report is also a key moment because it is the starting point for a kind of

self-evaluation in the country, to some extent independent of the visit by the experts and the publication of the country note. Country representatives confirmed that the drafting of the background report tends to fuel a critical debate about what the country is doing well or badly even before the experts provide their outside view (which is nevertheless the focal point of the review). The report often makes it possible to take note of endeavours undertaken in the field covered by the thematic review. It also covers other needs which may not necessarily be clearly evident, but are nevertheless effective in launching discussion and forcing policy-makers to take a stance.

The background report is, ultimately, an opportunity to bring together quantitative data. An enormous effort was made, at the time of the thematic review on adult learning, to construct a statistical base useful for a correct understanding of the mechanisms governing access to and participation by adults in training schemes. This was an extremely difficult part of the work, as reflected by the (low) level of completion of the proposed tables, which also highlights the need for qualitative field analysis. To date, few decisions in the field on education in general and adult learning in particular seem to be based on a solid quantitative evaluation of intentionally collected data, except for IALS or the programme for international student assessment (PISA). There are obviously exceptions, but there seems to be – especially as regards data on the nature and the extent of expenditure – a place for this type of investigation which is a counterpart to exchanges of experience.

3.2. The visits: seeing and understanding

This part of the review involves two visits of very different types and lengths. The pre-visit takes place some months prior to the visit, lasts only one day and involves only one member of the OECD Secretariat. Although it is an initial contact, experience has shown that it has a key role to play in reassuring the team in place in the country. The approach used for the thematic review is not, in practice, an audit-based approach. The pre-visit provides an opportunity

to say so, face-to-face, and to explain that this is a strategy of mutual aid, without penalties but with recommendations; this may mean that the team in place in the country is more disposed to accept it. It is a strategy of mutual aid because the team of reviewers is supposed to help the country to think about its own adult learning system and also to produce recommendations. In addition, exemplary practices of the country being reviewed may be useful to others. This acceptance of the method is essential if the thematic review is subsequently to generate new measures and reforms in the country.

The main purpose of the pre-visit is to help countries to draw up the programme of the visit. All key actors need to participate in the meetings with the OECD review team. As a minimum, the Ministries of Labour and Education need to be involved, although this could be extended to others such as the Ministries of Social Affairs, Agriculture, Defence, Industry or Health, if they are training ministries, and to those responsible for specific systems of qualifications. The visit, therefore, has to encompass the largest possible number of different partners (employers, trade unions, NGOs, training institutions, learners, etc.) Additional purposes of the pre-visit are to meet with background report authors, to clarify doubts and explain practical arrangements.

The team of experts is set up between the pre-visit and the visit. The standard team includes two members of the OECD Secretariat and three experts. The team of experts needs to include a balanced range of competences (economics, sociology, educational sciences, etc.), countries of origin and where possible, genders. This purpose of this balance and host of profiles is obviously to ensure good coverage of the themes and key approaches underpinning thinking about adult learning. An expert cannot be a national of the country visited: the main aim is specifically to invite experts who have an outside view and who, through their questions and interviews, are able to hold a virtual mirror in which the country's actors can see and evaluate themselves. In this way, the evaluation approach taken by the thematic review and its field visits, makes it possible to set ideas, and possibly reforms, in

motion in the country being reviewed. Selecting the experts is a key stage in ensuring coverage of all the issues of importance for the review country. Potential experts are always very interested in the thematic review and tend to agree to take part when first asked, provided that they are available.

Once the team of experts has been set up, the visit takes place. This lasts 5 to 10 working days, with between 50 and 100 meetings and/or visits, and leads ultimately to the drafting of a country note which is analytical in nature. The experts aim to get a general overview of how the system functions, its positive features, and potential pitfalls and to provide some recommendations. Four main themes, not necessarily completely independent, have been identified to help the experts to carry out their task of pinpointing good adult learning practices:

- (a) how the government, the social partners and the other actors can improve incentives and motivations for adults to learn;
- (b) how the fragmentation of adult learning can be reduced and replaced by an integrated approach to the provision of adult training;
- (c) how the quality of the learning (andragogy ⁽¹³⁾, etc.) and the variety of supply can be improved;
- (d) how the cohesion and the efficiency of adult learning policies can be improved.

These four themes have been chosen to enable a genuine review of the system – or embryonic system, depending on the case – in place. The four themes in practice cover the various levels involved. First is the national, or macro, level at which governments need to implement policies appropriate in terms of both incentives and motivations, but also to build up a system which is less fragmentary, more coherent, and more comprehensible to all users, adults as well as employers and trainers. There is then an intermediate level, that of institutions – public as well as private – which, in different ways and to different extents are depositories of adult learning. Finally, there is the micro level of individuals – and therefore learners – as well as all the small structures (microenterprises or small teaching units, for instance).

⁽¹³⁾ Adult pedagogy.

The field visits are essential to the thematic review's evaluation process since they make it possible to pinpoint measures that have been successful in one of the areas highlighted by the review's four main themes and/or at one of the levels identified. They also make it possible to measure the stage reached with thinking at field level and, in particular, to find out whether there is any consistency between what is being said at central level and what is happening or being said at local level. If there is no consistency on this latter point, the thematic review experts will pinpoint a lack of links between thinking and action, or even between decisions and action. This link is one of the prerequisites for the success of any policy.

The visits have to be correctly designed in terms of their length and their coverage of the issues to be examined. They should make it possible to flesh out, and therefore explain, what may emerge from data and/or analytical work. A field visit helps in particular to explain differences between differing, or even opposing, systems, or between choices or options which are apparently identical.

The meetings most systematically organised include an initial meeting, at the very beginning of the visit, with the authors of the background report, to provide the visiting experts with an opportunity to make sure that they have understood the content of this report. A meeting is also organised with adult learning statisticians, at the national statistical institute, as well as a meeting, often very long and lively, with grass-roots, university or other researchers. In the last meeting, the experts present the preliminary findings that will be included in the country note to the country representatives, who will validate it. The aim is for the visiting group's rapporteur to check that a major theme has not been neglected or that too much attention has not been paid to another theme.

The whole of the visit is therefore built around the idea of a continuous exchange with the country visited. However, the programme of visits may not be completely objective. Even if it is organised in conjunction with the country, it is

difficult to find out whether the range of experiments selected to provide the experts with an overview of what is actually happening is representative of the adult education and training universe. The teams of experts do not generally know about the country in advance. Their analysis is shaped by the visits that they are offered, which may lead to a slight bias. Along the same lines, it is also easier to see good rather than bad practices. It is very difficult in practice to gain access to experiments that are not working or are not very successful or to those that have not worked: countries are obviously more inclined to show off their successes than their failures. It is, nevertheless, of considerable value for the experts to view all initiatives and programmes. This allows them to compare successes and failures in their respective contexts and prevent other countries from making the same mistakes.

3.3. The country note and the comparative report: analysing

At the end of each visit, the rapporteur, who is one of the three experts who is not a member of the OECD Secretariat, draws up a country note. In contrast to the background report, which is intended simply to explain the state of play to the experts, the country note is intentionally more analytical and sets out recommendations in the various areas covered by the review's four themes. The drafting of the country note is a period of intensive exchanges with the review country. The rapporteur can use the contacts that he or she has made in the country to obtain further explanations of points that continue to be unclear and each version of the country note produced is sent to the country for comment ⁽¹⁴⁾.

Although the rapporteur has considerable latitude as to the way in which the experiments and programmes under way in the country are addressed and analysed, he or she also has to address the issue of adult learning from the point of view of the four themes specified in the terms

⁽¹⁴⁾ The country note also reports on the points of view of the other experts in the team, as a meeting is held every day, after the field visits, to discuss the validity of the programmes examined, the mechanisms identified and the contributions that they are making.

of reference and approved by the countries in the various OECD committees.

The country note is a document that is used widely in the country and there is little doubt, when it is published, that some countries use the OECD thematic review to breathe new life into the internal debate in their country. This dimension should not be underestimated in the evaluation of the thematic review of adult learning. An authority outside the country is often in a better position to exert pressure, in a neutral way, for a dossier that has been partially forgotten to come back to attention.

One of the thematic review's few problems emerged when the structure provided by the four themes mentioned in Section 3.2 had to be retained when drafting the country note. Even though these themes seemed to be natural and appropriate themes – and had been approved by the countries – they did not match what was actually happening. Some redundancies and issues that were new or insufficiently taken into account emerged. Each author in some cases found that they were overly restricted by the four themes decided in advance. It also proved difficult for the experts to agree on the theme best able to cover a particular issue. To cite only one example, the first theme examines how adults can be motivated to start to learn again and what resources are being used by the authorities to encourage this motivation; the thematic review showed in practice that this was a key dimension. It very quickly became evident that any measure encouraging learning, because it generates incentives for individuals, also generates motivation. It also emerged that almost all the issues dealt with during the visits could be grouped under this first theme. It became necessary,

therefore, to deal with the problems raised by the redundancy of some points of the questionnaire and to choose how to break down the themes to be addressed during the drafting of country notes and the comparative report.

Drafting the comparative report is the final stage of the first round of the thematic review and is the task of the OECD Secretariat. This exercise places the comparative aspect of the study on a concrete footing. The aim is to put everything that is of importance into perspective in order to understand, evaluate and advance adult learning programmes from the point of view of the four key themes dealt with in the country note and set out in the terms of reference. Comparing countries and explaining to everyone what each country is doing and how – or at least under what conditions – it is working is the study's second main contribution, after the country note. Finding out under what conditions and circumstances given practices may be considered good practices helps countries, in theory, to shape their own policy. The importance of comparative work in any evaluation approach goes without saying. In this respect, comparative evaluation is integral to the thematic review.

In summary, all the stages of the thematic review process are constructed so that they optimise the depth and quality of the review, with continuing exchanges, during preparations for the visit and during the visit itself, between the people responsible for the review in the country and the experts. These different stages are constructed so that they shed light on the issues and make it possible to take measures which are known to work under certain conditions. Chapter 4 looks at the measures taken by the countries after the thematic review.

4. A range of evaluation results

Evaluation of the results of the thematic review's analytical work is a difficult but necessary exercise. The aim is to try to evaluate the impact, or potential impact, on the countries and on decisions on adult learning in general. It is possible, even at an early stage of the review, to find out what uses countries are actually making of the OECD's thematic review in their national context. It seems that the review has contributed to the promotion of a diversity of approaches and good practices and has fuelled current international and national discussions on adult learning. In addition, public availability of all the documents prepared for the review provides international research and decision-makers with a valuable source of information on adult learning policies and practices in the OECD countries.

It is still too early to have a clear and concise view of the impact of the thematic review of adult learning on policy-making in the nine countries taking part in the review. The country notes have almost all been completed in previous years, as the visits took place, but the international comparative report has only recently been published (OECD, 2003). The exercise has just recently reached a conclusion and time is now needed before the impact of the thematic review can be measured.

At the time of writing this article, countries have benefited from the country-specific part of the

activity, with a detailed analysis of their own situation, without reference to the international comparative perspective, and many of them have already noted that this work has been very useful. They have, nevertheless, received a provisional version of the comparative report for comment. This report was not in a suitable state for wider distribution, nor has it been possible to analyse its findings and proposals in any depth in terms of its impact on countries and their policy choices. Following publication of this report, countries will need a further period to reflect on the comparative findings before any changes are proposed, either within the countries studied or, in the long term, possibly by the OECD Secretariat.

The following section describes the changes under way and the reforms adopted. It gives an overview of the types of recommendations made and clarifies the impact, whether direct or indirect, that the thematic review may have had.

4.1. Positive country responses

Nine countries took part in the thematic review. In all the countries, major reforms of adult learning systems were undertaken during or just after the exercise (Table 1).

Table 1: Recent reforms of adult learning policy

Country	Recent reforms (up to December 2002)
Canada Visit: January 2001 Country note: April 2002	Skills and learning agenda recognises the need to target adult learning in particular. <i>Knowledge matters: skills and learning for Canadians (2002)</i> targets post-secondary education, highly skilled workers and immigration. <i>Achieving excellence: investing in people, knowledge and opportunity (2002)</i> to step up investment in innovation and ensure a high level of research and development.
Denmark Visit: November 2000 Country note: September 2001	<i>Better education (2002)</i> is a new reform to improve the education system at all levels. For adults, this includes: more flexible vocational training, the reform of continuing vocational training programmes and the provision of more comprehensive and flexible programmes. Development of methods of assessing people's prior experience and real skills.

Finland Visit: February 2001 Country note: June 2001	A report prepared for the Ministry of Education and Science (2002) by the Adult Education and Training Committee put forward a strategy for adult education and training and key measures are planned up to 2010. It is proposed: – to improve the standard of education of adults with little training and to launch a programme of separate measures; – to strengthen the mechanisms governing adult vocational training, consolidate the financing system and improve the efficiency of networks of training organisations in the area of post-secondary vocational training for adults; – to finance, steer and formulate the objectives of popular adult education.
Norway Visit: March 2000 Country note: November 2000	Improvements to the Competence reform: A committee has been set up to locate models making it possible to cover people's subsistence needs during training leave; it has already decided not to use a model with centralised funding. There have been recent budget cuts in adult learning. Parliament has called for a new report on adult education.
Portugal Visit: March 2001 Country note: October 2001	ANEFA ^(a) has been integrated into the new Directorate-General for Vocational Training (which is responsible for young people and adults) within the Ministry of Education. One of the objectives is to provide adults with few skills with qualifications and/or to enable them to have their skills certified. A new law on vocational training is expected shortly.
Spain Visit: November 2001 Country note: December 2002	New law on qualifications and vocational training (May 2002). Plans to create a national qualification system integrating the different types of certification of vocational training and taking account of informal learning.
Sweden Visit: May 2000 Country note: September 2001	New bill on <i>Adult Learning and Adult Education (2002)</i> so that a better response, not just shaped by supply, can be found to individual needs: Creation of new institutions (CFL for distance learning, a national agency for advanced vocational training, reorganisation of the national education agency, an institute for the education of people with specific needs, recruitment grants, counselling and certification).
Switzerland Visit: December 1999 Country note: August 2000	A law had already been scheduled, and its outlines had been widely discussed at the time of the visit. In November 2002, the Berne government site announced that final consultation in Parliament was imminent. The introduction of the law is scheduled for 2004, or taking account of any delays, 2005.
UK Visit: February 2000 Country note: June 2001	Changes geared towards greater consistency and more public sector support for adult learning. Creation of the <i>Learning and Skills Council</i> in April 2001. The Sector Skills Councils, at local level, play a more important role. The overall aim is to rationalise the financing system and learning policies, and for employers' associations to improve the qualifications of workers in industry and business. Temporary discontinuation of the Individual Learning Accounts.
NB: The date given for the country note is the date on which the first version was sent to the country's authorities. Expressions in italics are free translations of programmes or policies with official names.	
^(a) ANEFA: National agency for adult education and training.	

Examination of Table 1 raises a number of questions: were these changes brought about by the OECD review? To what extent did the recommendations put forward by the teams of experts and the international comparison influence these

reforms? What other variables also played a part in recent developments in adult learning policy?

As this work includes several strands ranging from the collection of aggregate data to analysis, from consultation and exchanges of information

with countries to evaluation and comparisons of countries, the impact of this evaluation and its benefits for these countries varied greatly. Several comments can be made at this stage. They have been divided into general and more specific comments.

4.2. Raising awareness: adult learning is a key policy issue

The thematic review helped to generate an awareness of the importance of adult learning. Close cooperation with the delegates of the countries taking part and with those wanting to take part in future (in the second round), meetings with the main actors during the country visits, presentation of the main findings at national conferences, OECD Education and Employment committee meetings, International Labour Office (ILO) committee meetings and meetings within the various authorities of the EU, have helped to disseminate the results widely. A regular newsletter has also been created to distribute information on the activity. After the country visit and the receipt of the country note, most of the national coordinators drew up ad hoc reports in their own language for widespread public distribution, for members of Parliament or for working parties. These reports summarised the main findings and set out the recommendations and conclusions of the OECD teams of experts. Some included personal reports on their involvement in this work and the lessons that they had learned.

This dissemination work has played a key part in promoting adult learning in national education and training policy agendas. Several Ministries of Education of OECD countries told the teams of experts that they needed arguments to increase education expenditure and particularly expenditure on adult learning. There is a consensus that adult learning has been very under-funded until recently. Improving awareness is one of the main outcomes of the OECD's thematic review: making people aware that greater activity is taking place at national and international levels; that the OECD has launched a major research programme in this area; and making people aware of the problems and potential benefits that lay behind the need to increase adult learning opportunities. All this is

due to the thematic review and offers remarkable help for the countries.

This review also has to be seen in the wider context of the activities of the OECD Secretariat. Countries are constantly undertaking reforms of their education and training systems and the Secretariat is an integral part of this. The OECD has 30 members and information on, and analyses of, policies and programmes under way come from their institutions. As a result, an exercise of the type of the thematic review is an integral part of the movement to improve adult learning systems in a lifelong learning perspective. The Secretariat can play a central part in these national and international endeavours and is often the catalyst, if not the prime mover, in this area because it is able to collect and analyse information from a large number of countries.

Other international endeavours should also be mentioned as they play a key part in promoting lifelong learning. They are also central to this process and have had a major impact in this field. The European Commission's endeavours have, in this respect, been very important for its Member States. The process of consultation on the use of lifelong learning that it set in motion with the Memorandum of the same name and the subsequent *Communication Making a European area of lifelong learning a reality* have led to important definitions and proposals for action. As part of its endeavours for 'education for all', Unesco's Institute for Education has included lifelong learning on its agenda. The ILO has also set in motion a programme of work focusing on qualifications, knowledge and employability which has highlighted lifelong learning in general, and adult learning in particular. A large number of NGOs and associations are working on this issue as well. The OECD is embedded in this process of policy reform, on a par with countries themselves and other organisations.

4.3. High quality evaluation by outside experts may trigger action

The external evaluation process has two strands: a process of self-evaluation which includes the drafting of the background report and the organisation of the country visit (Chapter 3). This

provides the country with an opportunity to describe past successes and the stage reached with current programmes. Reforms are examined, questions discussed and future directions are clarified. It is then the turn of the OECD experts to evaluate the picture as a whole. The country is then able to take stock of the results of the evaluation in order to design its future policies.

The thematic review offers an independent evaluation which gives countries and decision-makers a chance to learn and to improve their policies. It is also an instrument through which government action can be reviewed, and a source of independent and valuable information for organisations not directly involved in the actual activity. It adds to knowledge of problems by providing a wider and independent overview and acting as a point of reference for experiments being run in areas other than the actual activity. However, internal monitoring and evaluation must also be a key part of the design and application of policies for adult learning. In practice, this is one of the most frequently noted shortcomings of the national policies implemented. External evaluation is a supplement to the work to be carried out internally and is in no way a substitute.

External evaluation may also offer the country an opportunity to set up methods to foster cooperation between institutions which may have different outlooks. This may help them to reach a consensus on solutions that enable each to achieve its own objectives. One of the thematic review's conclusions which the Ministries of Education, Labour and even Industry have had to face is that they need to work together with similar objectives and in full cooperation. In practice, some countries have already expressed concerns about this issue in their recent reforms. The Competence reform in Norway includes a strand on policies and practices entitled *Working together*. The aim is to coordinate the various levels of the institutions involved and this, in practice, is one of the main objectives of the reform.

4.4. Improved national and international cooperation

The most immediate practical consequence that can be seen in the countries taking part in the thematic review is the high level of cooperation

between institutions that participated. This is seen as one of the successes of the work. These exchanges between different potential partners took place at all stages of the work: during the preparation of the terms of reference, the background report, the visits to the countries, the country note and during the drafting of the international comparative report.

This means that the exercise is not just an evaluation by outside experts but also that the countries are protagonists of their own evaluation. In many cases, the national coordinators who accompanied the experts in the field said that they had learned a lot from the discussions that they attended. The process, by offering an opportunity for experts from different disciplines and different countries to visit the field, may well lead to questioning that is in some cases out of the ordinary and often disturbing, and may also generate fresh debate among the actors visited. Taking part in these exchanges and seeing the points of view emerging from them is highly educational, even for people *au fait* with the national debate on adult learning.

The review process also led to cooperation between various actors in the countries visited. In many cases, the Ministries of Education, Labour and others worked together to ensure that the thematic review was a success. These Ministries were represented almost systematically in the main meetings, so providing an opportunity for the OECD experts to raise questions about the general operation of the system and to think about questions of interest to all the main parties involved.

As regards cooperation between countries, the decision-makers in charge of adult learning and officials from the Ministries involved in the various countries exchanged points of view, opinions and practices and learnt from one another. The provision of a framework for thinking common to all the countries made it possible for them all to use and understand the documents prepared. The final comparative report brings together some of the comparisons offering the best lessons for adult learning policies and practices.

This cooperation in practice helped all the participants to gain a clearer understanding of the issues and to exchange information and data. The method, essentially comparative, helped those in charge to take important decisions as

regards adult learning, to think about their own approaches to the issues and to gain a picture of successful initiatives in other countries or in other regions of their own country. For instance, the thematic review gave the Ministries responsible for adult learning a chance to work together and to exchange both descriptive information and development prospects. It also paved the way for cooperation and consultation between those in charge of adult learning policy and other actors in the field.

4.5. Some general recommendations from the thematic review

Countries are looking into ways of improving existing adult learning systems or of devising programmes to step up adult skills and competences. Ministries of Education and Labour are looking for arguments to increase the financing of adult learning or to rationalise funding mechanisms and make them more efficient. Most governments, therefore, agree that intervention in the adult learning market is justified by arguments of efficiency and equity and have intervened in various ways, ranging from the formulation of general plans of action to a much more precise targeting of adults' needs in education and qualifications. These include adoption of laws defining the role of the institutions responsible for adult learning or regulating other practical issues concerning their own adult learning structures. The thematic review is a source of information on these various approaches and on the aims that they are pursuing. It also gives the point of view of the international experts appointed by the OECD.

All the countries involved in the first round of the thematic review are in the process of devising new policies and measures or are implementing fundamental reforms in the area of adult learning and lifelong education and training. Analysis of adult learning systems shows that no model is any better than any other as regards the delivery of adult learning from the point of view of either efficiency or equity, or in terms of achieving the objectives being pursued. Countries all have different objectives or different ranking of their priorities. The system is also rooted in different political and social structures which may have a

major impact on the way in which adult learning is organised. A large number of actors are involved in the design and supply of adult learning: different Ministries, the social partners, employers and individuals, among others. These actors play different roles in different countries.

Since there is no best model, governments may work in different ways. They may, for instance, try to increase adult participation in learning, or to achieve greater equity by providing learning opportunities for those who need them most. They may also try to improve adult learning for purposes of economic efficiency or to satisfy the current trend towards lifelong learning. There is, nevertheless, a consensus in all the countries that the government has a central role to play in setting objectives or priorities. This may depend, however, on the role assigned to each of the partners involved: the individual, the employer and the government as well as the political structures in place and the degree of intervention in the private market that the country is prepared to undertake.

There is also a fairly widely shared consensus that it is necessary to pinpoint who should provide different types of learning; this has financial implications. It would seem that the public sector concentrates on basic skills and second chance schools (primary and secondary levels). The Ministry of Labour tends to be more responsible for vocational training for the unemployed and, in some countries, it also provides assistance for in-company training; in others, however, this is organised by the social partners or left to market forces. In some countries, popular education is also within the public sector as is language learning, the new technologies or even learning for personal ends. This segmentation of resources and political responsibilities makes it difficult to gain an overall picture of adult learning. Moreover, different groups of the adult population do not all have equal access to learning opportunities.

A general recommendation of the thematic review is that countries should adopt an integrated approach to adult learning policies, including the following essential ingredients:

- (a) measures to make learning more attractive to adults and ensure that it covers the various needs identified. This may involve, among other things, teaching methods geared more

- towards adults, more flexible provision, outreach policies, guidance and counselling and recognition of prior experience;
- (b) measures to stimulate vocational training in enterprises, for workers, and for the unemployed. This includes entitlements to training leave, ensuring that skill reviews are available and the possibility of investment in training at enterprise level. The public employment services undoubtedly have a role to play here;
 - (c) financial incentives. As there is no consensus about the best way of financing adult education and training, countries can choose different financing models depending on the level of involvement of the various partners. The introduction of individual incentive mechanisms is one option. Offering entitlements to training leave or training hours during working time is another. Lastly, some countries offer subsidies to private providers or individuals for training purposes, or have introduced a learning tax;
 - (d) strategies to improve the quality of adult learning. These may focus on quality control and measurements of learning outcomes. Advances can be made, for instance, by improved monitoring of training schemes, improved collection of statistical data, development of certification systems, more direct involvement of institutions in performance evaluation and the development of systems to monitor how learners and trainees fare in the future. Research in this field is particularly important;
 - (e) taking a coordinated approach to adult learning by encouraging all the partners to work together is also crucial. Coordinating the work of the various actors may help to rationalise the few resources available and to pave the way for an efficient public expenditure management system. Partnerships are key tools in this area, as is a result-based approach. Countries are also trying out new methods to develop an exhaustive and integrated framework for adult learning policies. In contrast to the piecemeal approach seen in many countries, a holistic vision – including formal, informal and non-formal learning as well as general education, vocational training and in-company training – requires coordination.

4.6. Recommendations specific to countries reveal similar problems and issues

General recommendations on the introduction of instruments needed for a coherent policy of adult learning have arisen as a result of the research undertaken as part of the thematic review and the analysis of the good and bad practices examined during the field visits to the countries. The teams of experts learned a great deal from their observation of good and less good practices in the area of adult learning. This made it possible to draw up specific recommendations for each of the countries in the country notes prepared by the nine teams of experts. Examining these recommendations in conjunction with the analysis of recent developments in adult learning policies in the countries shows that the thematic review tackled issues that are crucial at present (Table 2). Most of the questions examined by the teams of experts are also discussed regularly in the countries and have fuelled discussions leading to the current reforms. It is very difficult, however, to establish clearly whether these reforms are a direct result of the OECD recommendations and the work conducted during the thematic review. They could also be the result of the cooperation between countries and the OECD Secretariat or could have arisen naturally as a purely national initiative. The truth probably lies somewhere between these two positions and depends on the country in question. In any case, formal documents and more informal conversations with the authorities of the participating countries have shown that the documents produced by the thematic review have had a fairly substantial impact on current developments.

One of the key factors to be stressed is that the countries are reflecting on similar issues concerning adult learning. Although they were visited by different teams and the country notes were drafted by different rapporteurs, many of the issues tackled are similar: the need to take a more coordinated approach; the recognition of non-formal and informal learning; the need to improve financial incentives to encourage adults to learn; and the need for more systematic evaluation and better information and guidance to help adults find their way in the world of learning. This list can be supplemented by the need to diversify

It has to be borne in mind, in the context of thinking about ways of evaluating the specific contributions made by the thematic review, that the way in which it is organised is not neutral as regards the nature of the recommendations made. As mentioned above, each country was visited by a different team and different rapporteurs. This may lead to a situation in which some themes receive more attention in one country than in another, depending on the profile of the experts, for instance, or the type of sites visited. The fact that no recommendation is given for a particular issue may simply be due to the fact that no expert expressed an opinion on the need to strengthen this or that aspect. It may also be due to the fact that the team of experts did not visit experiments or practices which led them to discuss this issue during the visit. Lastly, and it is undoubtedly for this reason that it is difficult to evaluate the thematic review exercise, a recommendation may not have been made because the country had already taken the most appropriate steps to resolve the question.

Pinpointing the role played by the thematic review in these developments is easier in some

countries than in others. In some cases, OECD documents were included in the background documentation used for discussions and preparation of reforms. This was true of Canada, where OECD documents were clearly evident. In Denmark, the thematic review recommendations on the recognition of informal and non-formal learning were followed by the development of 'better education' programmes. In Finland, where a new programme for adults with few qualifications is to be launched shortly, the thematic review is felt to have had an impact. In Portugal, a new reform of the institutions working in the area of adult learning drew on the OECD's work and recommendations. In the other countries, even where there is no explicit reference, the reforms are also moving in the direction of the OECD recommendations: taking a more demand-oriented training approach in Sweden and improving the recognition of informal and non-formal learning as well as improving evaluation strategies in Norway. However, it should be noted that countries were not really able to clarify the exact role played by the thematic review in these recent developments when they were consulted.

5. Comments and prospects – an exercise worth pursuing

Evaluation needs to be an integral part of adult learning systems as it helps to improve the decision-making process. Evaluating results may help decision-makers to choose the adult learning programmes best geared to specific objectives. It is then essential for all the partners involved, from learners to investors, to be able to understand the outcomes of the evaluation and gain a picture of the issues involved in adult education and training. To obtain a consensus, keep up the commitment of the actors and make policies viable in the long term, it is vital to know how effective and useful the measures taken actually are and to assist their implementation. The thematic review is integral to such endeavours.

The only exercise that could genuinely clarify the relevance of the thematic review in terms of its impact on the measures and reforms under way in the countries would be to analyse the extent to which the countries have adopted the recommendations suggested to them. This is a real challenge as policy decisions do not always have discernible or well-documented justifications. Many other variables play a part, moreover, in decision-making mechanisms and it is impossible to measure their respective contributions. Particular developments may have taken place in the country well after the thematic review and may have to do with the ideas put forward in the country notes. It may also be that the enthusiasm generated and the synergies developed between the different partners involved have had an impact. This article has tried to relate the thematic review recommendations to recent developments. Even though it is still too early to pinpoint a real cause and effect relationship in the countries concerned, there is a broad agreement that the exercise is of genuine relevance.

It is already possible to draw some conclusions about the value of the exercise. Whether directly or indirectly, it has had an impact on national policy developments in adult learning. The benefits include the fact that this experiment has provided international research and the community of decision-makers with a database on adult learning policies and practices. This can

be consulted on the Internet. The information gathered and the documents produced offer a very broad picture of the different approaches that countries are currently taking or studying to improve their system. The review provides research information that may help to improve adult learning policies, not just in the review countries but in other countries as well. The documents produced will be points of reference, especially bearing in mind the fact that the OECD itself is often a point of reference. This could be one of the main contributions of the thematic review. The way in which the international comparative report, published in February 2003 (OECD, 2003), is greeted will give some idea of the interest and relevance of the questions addressed.

More generally, since the OECD Secretariat is, by definition, in a better position to draw up comparative reports than a single country, it is often commissioned to work in fields where it is necessary to collect quantitative data, put together more qualitative information, locate best practices and make international comparisons. This was the starting point for the thematic review. The interest expressed by nine countries in participating in a second round indicates the relevance and usefulness of the work undertaken in the first round for the countries reviewed, and for observers who are not directly involved.

Evaluation may take different forms and many public programmes are currently being evaluated to determine the pertinence of investment or to steer and assist reforms. The OECD's thematic review, while evaluating the national policies and practices of the nine participant countries in an international perspective, also sheds light on those of other countries. When complete, the exercise will be able to highlight examples of good practice and analyse what is working and why it is working. Countries are also provided with a very rich source of descriptive information. Nevertheless, this kind of evaluation might not satisfy countries expecting more in-depth evaluation. This international evaluation using common themes does not, however, disregard more

detailed, more contextualised and more specific national evaluations. The OECD Secretariat may recommend such evaluations but undoubtedly cannot initiate them. One of the main findings of the review is that countries themselves should organise more systematic evaluation of the programmes that they introduce.

There are various ways of building on the review, the most obvious being to increase the number of countries involved. This is under way: a second wave of countries has been organised for 2003 and 2004. Nine countries are to take part: Austria, Germany, Hungary, Korea, Mexico, the Netherlands, Poland, the UK and the US. A second round of countries offers an opportunity to check and refine the analyses of the first round. The aim is to evaluate, in different political, social, economic and cultural contexts, some of the hypotheses made in the first round of countries. Making the group of countries reviewed less comparable – extension to Eastern Europe and to countries with a dual system of training, for instance – will obviously help people to understand what makes the thematic review such an interesting evaluation tool.

A second round is also a very good opportunity to organise thinking around recurring themes common to all the countries of the first round. Analysis work could also be limited, so that it can be in greater depth, to particular sub-groups of

the population, could look in detail at some aspects, and could return to and/or take up again some analyses in the light of the lessons learnt from the new countries. The first round, for instance, demonstrated how important it is to pay sustained attention to individuals with low skills.

It will also provide an opportunity for new questions to emerge and, retrospectively, to address them in the context of the first-round countries. A second round could also be envisaged in the countries taking part in the first round. The aim would then be to study what has happened, to open up new discussions and/or take up any points left in abeyance during the first round – new regions, new population groups – in order to conduct a focused review. This is the strategy that Germany, the Netherlands and the US have decided to take.

Finally, another possible measure would be to return to the field after a period of time – four to five years for instance – to assess the actual impact of the review and its recommendations on adult learning, its programmes and its actors and users. This is one of the prerequisites for assessing the real relevance of a thematic review and its ability to generate the necessary changes in adult learning in a country. A final stage needs to be conducted within a reasonable period if the evaluation tool represented by a thematic review is to be evaluated.

List of abbreviations

ANEFA	National Agency for Adult Education and Training
ELFS	European labour force surveys
IALS	International adult literacy survey
ILO	International Labour Office
NGOs	Non-governmental organisations

References

- Human Resources Development Canada. *Knowledge matters, skills and learning for Canadians*. Quebec: HRDC, 2002. Available from Internet: <http://www.innovationstrategy.gc.ca> [cited 2.9.2003].
- Ministry of Education and Research Norway. *Funding for subsistence in connection with leaves of absence for educational purposes*. Oslo: AAD, 2001 (Official Norwegian Report NOU 2001, No 25).
- Moser, C. *Improving literacy and numeracy: a fresh start*. London: Department for Education and Employment, 1999.
- OECD. *Beyond rhetoric: adult learning policies and practices*. International comparative report of the thematic review of adult learning. Paris: OECD, 2003. Available from Internet: www.oecd.org/edu/adultlearning [cited 2.9.2003].
- OECD; Statistics Canada. *Literacy in the information age*. Final report of the International Adult Literacy Survey (IALS). Paris: OECD, 2000.
- Undervisnings Ministeriet Denmark. *Better education*. Copenhagen: UVM, 2002. Available from Internet: <http://pub.uvm.dk/2002/better2> [cited 2.9.2003].

Measurement and evaluation of competence

Gerald A. Straka

Abstract

Development, measurement and evaluation of competence has become an important issue in education and training. Various factors contribute to that:

- (a) the shift of focus in education from input to output, stimulated by standard-based assessment; accountability systems;*
- (b) international comparisons of school system achievement;*
- (c) the transition from subject to literacy orientation in education;*
- (d) recognition of learning in non-formal and informal settings, the vision of lifelong learning as a prominent goal of the European Union (EU).*

Nevertheless, electronic literature research into vocational education and training (VET) using, in combination, the keywords 'measurement', 'evaluation' and 'competence' generated a negligible number of references. Better results were obtained using 'competence' and 'evaluation'. This may indicate that the link between measurement, evaluation and competence is still absent in European VET research.

Measurement requires deciding what is to be measured. In order to provide specification, a general diagnostic framework is introduced. It consists of three properly differentiated levels: external conditions (e.g. situations, products), actual episodes realised by an individual (e.g. behaviour, cognitive operations, individually created information, motivation), and personal internal conditions (e.g. knowledge, skills, motives). Within this conceptualisation measurement faces the problem that knowledge, skills, and cognitive operations are not visible to outsiders. The only way to achieve insight is by observation of specified external conditions and/or behaviour (e.g. realised behaviour, changed situations, or created products). From this observation, specified/defined elements of the internal conditions (e.g. knowledge, skills) are inferred. The relations between the observable and non-observable are established through interpretation rules and hypotheses.

Evaluation, in this context, is judging the observed competences against defined benchmarks. Such benchmarks may be measured knowledge, skills, actions, or performance of other persons (norm-referenced); or they may be theoretically specified types and levels of knowledge, skills, actions, or performances (criterion-referenced). Evaluation and measurement themselves are subsumed under the term assessment.

Against the background of this general diagnostic framework, selected competence definitions in the EU are analysed. Most of them bridge all the three levels of the model. Such a broad and multilevel concept of competence might generate more misunderstanding than understanding in public and scientific discussions. Therefore, two recommendations are made on how to define and assess competence:

- (a) an accurate description of the tasks and requirements (= external conditions);*
- (b) a specification/characterisation of the psychic attributes a person should possess or that has built up in a specific occupational domain.*

Selected procedures of assessing competence in the EU are introduced:

- (a) the bilan de compétences (France);*
- (b) the national vocational qualifications (NVQ) (UK);*
- (c) dimensions of action competence in the German dual system;*
- (d) assessing competences at work (the Netherlands);*
- (e) realkompetanse (Norway);*
- (f) recreational activities (Finland);*
- (g) competence evaluation in continuing IT-training (Germany).*

The analysis of these conceptions using interrelated assessment quality criteria (validity, reliability, objectivity, fairness and usability) revealed that, to date, empirically grounded findings about fulfilling these criteria are scarce. Furthermore, methodological considerations indicate that self and other observation, even guided by criteria and realised with more than one external assessor, produce many errors.

The validity of the approaches varies considerably when it comes to diagnosing occupational competence and occupational success in general. The discussions often focus on the format of the tasks (closed versus open-ended). In this respect, the advantages of performance-based assessment through open-ended tasks requiring complex skills (i.e. involving a significant number of decisions) should not be overestimated. The problem is that the increased cost of evaluation is not justified by the increase in external validity. In addition, there is evidence that the key is not the format but rather the content requirements of the task.

Concerning the timing of assessments (e.g. continuously, sequential or punctual) valid evidence can only be obtained through systematic empirical investigations of assessment procedures in relation to concepts of competence development.

In addition, there are methodological reasons to supplement self- and peer assessment with standard oriented assessment and accountability considerations. Assessment should be done with regard to criteria, such as those proposed by the American Educational Research Association (AERA).

On the basis of these findings, the following recommendations are given to correspond with the EU goals of transparency and mobility:

- (a) initiating a detailed and summarising review of the diverse practices of competence assessment in the EU, focusing on the methodological dimension;*
- (b) promoting empirical investigation about the measurement quality of selected and prototypical assessment procedures practised in the EU;*
- (c) activating conceptual and empirical research on how to define and validate competence and its development in VET;*
- (d) advocating a VET-PISA in selected occupations or sectors under the patronage of the EU.*

Table of contents

1. Introduction	267
2. Measurement and evaluation	270
2.1. A general model of measurement	270
2.2. Evaluation	272
3. Competence	274
3.1. General competence categorisations	274
3.2. VET competence definitions	275
4. Procedures of assessing competence	277
4.1. The bilan de compétences (France)	278
4.2. The national vocational qualifications (NVQ) (UK)	279
4.3. Assessing action competence in Germany	281
4.4. Social and methodical dimensions of competence (DaimlerChrysler)	283
4.5. Competence at work: a case in the Netherlands	284
4.6. The realkompetanse project (Norway)	285
4.7. Recreational activities (Finland)	287
4.8. New ways of assessing competences in information technology (IT) in initial and continuing VET in Germany	288
4.9. Analysis and evaluation	290
5. When should competence be assessed and by whom?	294
6. Summary and conclusion	298
List of abbreviations	301
Annex 1: Sample overall training plan (Bundesanzeiger, 1998, pp. 8-9)	302
Annex 2: Sample of skeleton school syllabus (Bundesanzeiger, 1998, p. 18)	303
Annex 3: Unit 3 – sell financial products and services (NVQ Level 3, n.d.)	304
References	307

List of tables and figures

Tables

Table 1:	Comparisons of concepts	275
Table 2:	Validity of customary approaches	292
Table 3:	Performance-based assessment versus 'classical' approaches	293

Figures

Figure 1:	General diagnostic framework	271
Figure 2:	The structure of a NVQ	279
Figure 3:	Overview of the structural features of the dual system	281
Figure 4:	Excerpt from an observation and evaluation sheet for client-counselling of bank clerks	283
Figure 5:	Rating scale for the key competence 'ability to cooperate'	284
Figure 6:	Competence card from the workplace	286
Figure 7:	Example page of the recreational activity study book	288
Figure 8:	Structure of the IT-further education system	289
Figure 9:	From a practice to a reference project – an example for a network administrator	289
Figure 10:	General model of economic processes, a basis of business administration	295

1. Introduction

Competence and its evaluation have become an important issue in both the theory and practice of general education around the globe. This trend is not restricted to vocational education and training (VET). Several triggers can be identified of which two deserve emphasis:

- (a) a shift from input to output orientation in education through implementation in the US of a standard-based assessment and accountability system from the 1990s onward (Linn, 2000);
- (b) large-scale international achievement comparisons of school systems in different domains such as the *Third international mathematics and science study* (Baumert et al., 1997) ⁽¹⁾, or the *Programme for international student assessment* (Deutsches PISA-Konsortium, 2001) ⁽²⁾. Compared with former studies of this type, such as the first (1964) and second (1980-82) international mathematics studies, much more attention was given to the results of PISA 2000. Because PISA was replicated in 2003 with another focus – results will be published in 2004 – and will again be repeated in 2006 it is expected that the public awareness and discussion of measurement and evaluation of competence will continue.

TIMSS and PISA attest to a substantial shift in the generic aims of education and training. The focus of TIMSS was the contents of the mathematics and science education in schools of the participating nations. In contrast to this 'world curriculum approach' for the two subjects, the theoretical foundation of PISA is a 'literacy concept' not only in reading but in science and mathematics too. This concept additionally takes into account that the acquisition of literacy occurs not only during formal instruction but also in informal school settings, out of school, and

lifelong. 'PISA assesses the ability to complete tasks relating to real life, depending on a broad understanding of key concepts, rather than assessing the possession of specific knowledge' (OECD, 2001, p. 19). Such a perspective has an impact on the assigned situations and tasks as well as on the measurement models chosen.

In the US there was from the 1990s substantial pressure on the development and use of 'new' approaches to assessment under different labels such as alternative assessment, authentic assessment, direct assessment, or performance-based assessment (Linn and Gronlund, 2000). Against this background Lorrie Shephard ⁽³⁾ (2000) analyses and compares the curriculum, the learning and the measurement theories dominant in the 20th and at the beginning 21st century. At first, curriculum was guided by social efficiency. The concept of an innate, unitary, fixed IQ, combined with associationist behavioural learning theories led to sorting of students by ability; this ability was measured through objective achievement tests. Now, the curriculum is starting to be student-, problem- and authenticity-centred. It is in line with cognitive-constructivist learning theories. It uses formative assessment, addressing both outcomes and process of learning, and such assessments are used to evaluate both student learning and teaching. These findings suggest that assessment is interconnected with the curriculum concept and its learning theoretical basis.

Almost at the same time, two committees of the Commission on Behavioural and Social Sciences and Education of the National Research Council (NRC) of the US summarised the state of research on the science of learning and linked it with actual practice in the classroom in *How people learn* (Bransford et al., 2000). This report was the trigger for the Committee on the Founda-

⁽¹⁾ Under the auspices of the International Association for the Educational of Educational Achievement (IEA). The sample of TIMSS III focussing on 12th or 13th year of schooling included VET students in Germany. Their performances were analysed in relation to occupations (Watermann and Baumert, 2000).

⁽²⁾ Launched by the OECD in 2001.

⁽³⁾ Presidential address *The role of assessment in a learning culture*, annual meeting of the American Educational Research Association (AERA).

tions of Assessment to review and synthesise advances in the cognitive sciences and measurement as well as exploring their implications for improving educational assessment. The report – *Knowing what students know* – ‘addresses assessments with the three purposes: to assist learning, to measure individual achievement, and to evaluate programs. The conclusion is that one type of assessment does not fit all purposes’ (NRC, 2001, p. 1). As a consequence, measurement and evaluation of competence have to take in account their purposes, the concepts of learning and development.

In European VET, analogous developments took place – although mostly without explicit reference to the discussion in the US – of which examples are:

- (a) in 1987, in Germany ‘action orientation’ (*Handlungsorientierung*) was designated as the central objective of in-company training within the dual VET system. The school side adopted this objective in 1991 and, since 1996, the mission to generate ‘action competence’ (*Handlungskompetenz*) is the central pedagogical task of the schools in the dual system (KMK, 1996/2000). Linked with this development, ‘new’ approaches to assessment and examinations are requested, tested, or introduced;
- (b) independently from the German development, the national vocational qualifications (NVQ) were introduced in England and Wales in 1987 with ‘a very particular approach to defining and measuring competencies’ (Wolf, 1998, p. 207). In contrast to Germany, the NVQ concept explicitly refers to the competence-based training and assessment of the US (Wolf, 1995);
- (c) in 1991, *the bilan de compétences* was legalised and introduced in France, focusing on the competences acquired beyond formal schooling, during working life.

According to a feasibility study titled *Research scope for vocational education in the framework of COST social sciences*, transferability, flexibility and mobility should become the targets of VET in the future (Achthenhagen et al., 1995). These three targets are to be seen as heavily interrelated with the assessment of VET outcomes (Achthenhagen and Thang, 2002).

The conclusions of the European Council held in Lisbon in March 2000 mark a shift in the focus

of education policy towards lifelong learning, which ‘is no longer just one aspect of education and training; it must become the guiding principle for provision and participations across the full continuum of learning contexts’ (European Commission, 2000, p. 3). Lifelong learning is seen as the common umbrella under which all kinds of teaching and learning should be gathered and it calls for a fundamentally new approach to education and training.

In order to take action on lifelong learning, six key messages have been formulated by the European Commission with the purpose of carrying out a wide-ranging consultation on priority issues:

- (a) new basic skills for all;
- (b) more investment in human resources;
- (c) innovation in teaching and learning;
- (d) valuing learning;
- (e) rethinking guidance and counselling;
- (f) bringing learning closer to home.

The objective of valuing learning is to ‘improve significantly the ways in which learning participation and outcomes are understood and appreciated, particularly non-formal and informal learning’ (European Commission, 2000, p. 15). Regardless of the type of learner, innovative forms of certifying non-formal learning are considered important; absolutely essential is the development of high-quality systems for the accreditation of prior and experiential learning (APEL) (European Commission, 2001, p. 17).

These examples indicate interwoven trends which may be characterised as follows:

- (a) a change from input to output orientation in education;
- (b) a shift from the discipline or subject approach to competence and competence-based authentic measurement in education;
- (c) an increasing attention to the measurement and evaluation of competences acquired outside of educational settings;
- (d) methods of measurement and evaluation (= assessment) are regarded as interrelated with concepts of learning, development, education and training.

Do these developments since the 1990s have visible effects on VET at the EU level? A search in the Leonardo da Vinci project database for 1995-99 and 2000-06 with the combined keywords ‘measurement’, ‘evaluation’ and ‘competence’

yielded no result. Using other combinations, for example, 'assessment' and 'competence', 'diagnosis' and 'competence', 'certificate' and 'competence', only two project titles have been found. The search in Cedefop's VET-bib database with 27000 references using the keyword 'competence' found more than 800 in the titles, the keyword 'evaluation' more than 900, and 'measurement' more than 10, but the combination 'measurement', 'evaluation' and 'competence' gave no reference. The results with the two keywords ranged from 0 to 21 hits. These findings may be interpreted in several ways. The keywords may not be properly inserted in the database, or measurement in combination with

competence and evaluation is still absent in VET research at European level. In consequence, this contribution proposes to stimulate reflections on the issue by:

- (a) introducing a general diagnostic framework as a basis for measurement and evaluation;
- (b) selecting some competence concepts and analysing them against the background of this framework;
- (c) describing some assessment concepts in practice in selected EU Member States and analysing and evaluating them from a methodological angle;
- (d) outlining some implications for policy and practice.

2. Measurement and evaluation

The question of ‘how much’ refers to measurement, whereas the question of ‘how well does the individual perform’ refers to the evaluation of something observed and quantified against reference criteria. Both concepts will be developed in the following paragraphs.

2.1. A general model of measurement

Measurement as a systematic mapping ⁽⁴⁾ of psychological attributes and actions has to cope with the problem that parts of them are invisible. They have to be made visible for measurement and diagnosis. To locate and structure the observable and non-observable parts, a general diagnostic framework will be introduced (Figure 1). It differentiates three levels: the external conditions, the actual episodes, and the internal conditions as described below (Gagné, 1977; Klauer, 1973; Straka and Macke, 2002):

- (a) from the perspective of an acting or learning individual, a task, a situation, or a request (1 in Figure 1) and the created outcome (solution or product) (2) are elements of the visible socioculturally shaped external conditions;
- (b) the actual episodes operated by the individual are performed with the individual's actions. For their measurement it is crucial to differentiate between the observable and non-observable components of actions. The observable component of an action is the behaviour (3). The non-observable part of an action is the experience, which may be distinguished between cognitive (e.g. planning, controlling), motivational (e.g. achievement orientation) and emotional (e.g. fear, joy) aspects, all of them being part of episodes. It may be asked whether an action can be

realised independently of a context. The answer is no. Actions always need points of reference, such as mentally constructed representations of goals, intended outcome, pathways to the goal and criteria for bringing the process to an end. From the perspective of the individual these internal representations are actual information created by the individual and shaped by her/his prior knowledge or cultural perspectives ⁽⁵⁾ (Shephard, 2001; Bransford et al., 2000) as part of the internal conditions. It is the acting individual who is (re)constructing the situation and analysing it. From the individual's perspective a situation or a request is in itself part of her/his environment and, therefore, a ‘pure physical matter’ in the format of characters or signals (Straka and Macke, 2002, 2003).

It is important to point out that action and information are two inseparable sides of the same coin (4); they are differentiated only for analytical reasons. An example is the revised taxonomy of educational objectives (Bloom, 1956) which consists now of the ‘knowledge’ and the ‘cognitive process’ dimensions, both further subdivided and combined in a two-dimensional matrix (Anderson and Krathwohl et al., 2001). The components factual, conceptual, procedural, and meta-cognitive of the taxonomy's knowledge dimension correspond to types of information in the introduced general diagnostic framework. The cognitive processes of the taxonomy, such as remembering, understanding, applying, analysing, evaluating, and creating, name types of cognitive actions. The combination of these two dimensions and their components with the matrix-concept indicates that neither the information nor the action dimensions can exist on their own. Action and information are transient.

⁽⁴⁾ It should be noted that there are different qualities of measurement like nominal, ordinal, interval and ratio measurement level, which enable different types of conclusion and data processing (e.g. Kerlinger, 1973).

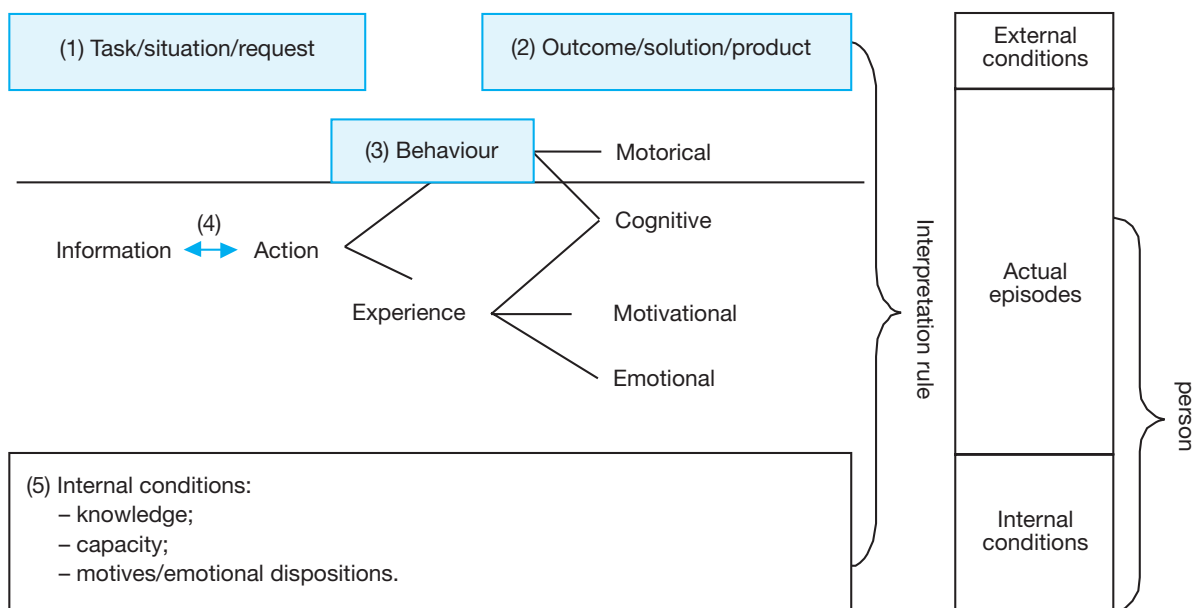
⁽⁵⁾ In using a hammer, the hammer is not in our brain but information about its composition, made of use, functions, etc.

They are present only in the moment of execution and immediately afterwards they belong to the past. Information as the result of an action stays only for fractions of a second in our working memory. It is not possible to rewind the interplay between action and information and to show it again like a movie. In order to keep on being able to act we have to rely on our knowledge and capability (located in our memory). Attributes such as knowledge, capability, motives,

emotional dispositions, and values are subsumed under the concept of internal conditions (5). In specific configurations they characterise an expert for a specific domain ⁽⁶⁾.

These considerations about the individual's interaction between the socioculturally shaped external conditions (environment), the actual episodes and the internal conditions as prerequisites and results of such a process, are put together in the following diagnostic framework:

Figure 1: General diagnostic framework



Source: G. A. Straka

This framework brings positive and negative implications for observation and measurement. The positive one is that outsiders can directly observe and measure the created and/or communicated outcome, the solutions and the result (2),

the task, situation, and the request (1), the difference between (2) and (1), and the behaviour (3) as well. All of them are performance elements which constitute the focus of performance assessment. The negative implication is that it is impossible for

⁽⁶⁾ 'Experts, regardless of the field, always draw on a richly structured information base; they are not just 'good thinkers' or 'smart people'. The ability to plan a task, to notice patterns, to generate reasonable arguments and explanations, and to draw analogies to other problems are all more *closely intertwined with factual knowledge than was once believed*' (Bransford et al., 2000, italics by G. A. Straka).

externals to observe, measure and evaluate directly the interplay between action and information and the internal conditions. The only way is to conclude from the observable, the non-observable. Considering that ‘observation [...] is always observation in the frame of theories’ (Popper, 1989, p. 31), theories, or at least hypotheses, about internal conditions and their development – like the pathways from novice to expert in a specific domain – are necessary. So are rules specifying the relationship between observation and theoretical constructs (indicated by the term ‘interpretation rule’ in the figure above; see also Shepard, 2000).

Summarising, assessment is always a process of reasoning from evidence. The results of such a procedure are only estimates of what a person knows and can do. ‘Every assessment, regardless of its purpose, rests on three pillars: a model of how students represent knowledge and develop competence in a subject domain (cognitions = internal conditions, actual individual operations), tasks or situations that allow one to observe students’ performance (external conditions), and an interpretation method for drawing inferences from the performance evidence thus obtained (interpretation rule). These three elements: cognition, observation, and interpretation rule constitute the angles of the assessment triangle’ (NRC, 2001, p. 2, italics by G. A. Straka).

2.2. Evaluation

In the context of individual environment interaction, evaluation judges measured competences against a defined benchmark (Straka, 1974). In this context, two approaches are distinguished: norm-referenced and criterion-referenced evaluation (Linn and Gronlund, 2000). In the case of norm-referenced evaluation, the measured competence is interpreted and judged in terms of the individual’s position relative to some known group. Criterion-referenced evaluation ⁽⁷⁾ is interpreting and judging the measured competence in terms of a clearly defined and delimited domain.

The two approaches have common and different characteristics. Both require a specification of the achievement domain to be mastered, a relevant and representative sample of tasks or test items. They use the same types of tasks and qualities of goodness (e.g. validity, reliability) for judging them. The differences – to some degree a matter of emphasis – are:

- (a) norm-referenced evaluation typically covers a large domain of requirements with a few tasks used to measure mastery, emphasises discrimination among individuals, favours tasks of average difficulty, omits very easy and very hard tasks, and requires a clearly defined group of persons for interpretations;
- (b) criterion-referenced evaluation focuses on a delimited domain of requirements with a relatively large number of tasks used to measure mastery, emphasises what requirements the individual can or cannot perform, matches task difficulty of requirements, and demands a clearly defined and delimited achievement domain (Linn and Gronlund, 2000, p. 45).

These different orientations of the two approaches have distinct impacts on the statistical measurement model. The norm-referenced assessment or the classical model is bonded with the normal curve whereas criterion referenced assessment is aligned with probabilistic models focussing on degrees of mastering a requirement and not related to the performance of other persons (Embretson and Reise, 2000). Criteria may be international, national, regional, institutional, individual standards such as personal goals, philosophies of companies (Section 4.4), state syllabuses, occupational profiles or NVQs (Sections 4.2 and 4.3). The core question remains: are these criteria concrete enough (operationalised) that decisions can be made on their basis. Since PISA there are some doubts – especially in Germany – indicated by the tendency to define national standards for subjects and age-groups. In the US, different content standards for school subjects or school types are already worked out (Linn and Gronlund, 2000, pp. 531-533). However, standards are varied. The reasons for this variety may be the specifics of

(7) Other terms for this type of evaluation are: standards-based, objective referenced, content referenced, domain referenced, and universe referenced.

the domains but also broad or diverse interpretations about what standards should look like (Bulmahn et al., 2003).

In the same way as the action is inseparably bonded to information in the course of the indi-

vidual's actual processing, the differentiation between measurement and evaluation is an analytical one. For that case the term 'assessment' is introduced and subsequently used, by subsuming measurement and evaluation.

3. Competence

The term competence has become increasingly popular since the 1980s in theory, practice and in VET political discourse. However, increasing use also led to an expanding meaning which will be illustrated by the reviews of Weinert (2001) and Eraut (1994). We will complement it by an analysis of some competence definitions used in VET against the background of the general diagnostic framework.

3.1. General competence categorisations

Weinert (2001, p. 45) introduces competence in general 'as a roughly specialised system of abilities, proficiencies, or skills that are necessary or sufficient to reach a specific goal' and he works out seven approaches by which competence has been defined, described, or interpreted:

- (a) general cognitive competences as cognitive abilities and skills (e.g. intelligence);
- (b) specialised cognitive competence in a particular domain (i.e. chess or piano playing);
- (c) the competence-performance model used by the linguist Noam Chomsky (1980) differentiating linguistic ability (= competence) enabling creation of an infinite variety of novel, grammatically correct sentences (= performance);
- (d) modifications of the competence-performance model which assume that the relationship between competence and performance is moderated by other variables such as cognitive style, familiarity with the requirements, and other personal variables (i.e. conceptual, procedural, performance competence);
- (e) cognitive competences and motivational action tendencies in order to realise an effective interaction of the individual with her/his environment; i.e. competence is a motivational concept combined with achieved capacity;
- (f) objective and subjective competence concepts distinguishing between performance and performance dispositions that can be measured with standardised scales and tests (objective) and subjective assessment of performance-relevant abilities and skills needed to master tasks;

- (g) action competence including all those cognitive, motivational, and social prerequisites necessary and/or available for successful action and learning (i.e. general problem-solving ability, critical thinking skills, domain-general and domain-specific knowledge, realistic, positive self-confidence, and social competences).

Eraut argues that competence is not a descriptive but a normative concept. 'In examining research into competence, we need to ask not only how competence is defined in general, but how it is defined in a particular situation, i.e. how these normative agreements are constructed' (Eraut, 1994, p. 169). Referring to Norris (1991) he distinguishes three main research traditions in post-war research in this field:

- (a) the behaviouristic tradition focussing more on training than on qualifications (competence-based training, e.g. in teacher education [Grant et al., 1979] widely spread in North America and heavily based on previous work in Canada [Ministry of Education, 1983]);
- (b) the generic tradition, rooted mainly in management education, widely used in Britain, for example Boyatzis (1982), listing 12 competences; e.g. concern with impact, self-confidence, or perceptual objectivity, and differentiating superior from average managers;
- (c) the cognitive competence tradition, most clearly articulated in Chomsky's theory of linguistics, which was adopted from research under the aegis of the UK Department of Employment and which is one of the rationales for the concept of national vocational qualifications (NVQ) in the UK.

Based on these two categorising reviews it can be concluded that the term 'competence' has the function of an umbrella for divergent research strands in human capacity development and its assessment. The umbrellas themselves differ in extent and differentiation with some overlap. Therefore, an additional approach will be chosen by introducing some definitions from the field of VET and analysing them against the background of the general diagnostic framework in order to draw some implications for measurement and evaluation of competence.

3.2. VET competence definitions

To give a flavour of how the competence concept in VET is used in different countries or institutions, definitions from EU level as well as from France, the UK and Germany will be introduced and discussed.

In the glossary prepared by the European Commission for the communication *Making a European area of lifelong learning a reality* (European Commission, 2001, p. 31) competence is defined as ‘the capacity to use effectively experience, knowledge and qualifications’.

Cedefop (Björnavold and Tissot, 2000, p. 208) define competence ‘the “proven/demonstrated” – and individual – capacity to use know-how, skills, qualifications or knowledge in order to meet usual – and changing – occupational situations and requirements’.

In France, Gilbert (1998, p. 9) defines competence as an entity of theoretical knowledge, ability, application knowledge, behaviour and motivation structured in mastering a specific situation.

Within the system of the national and Scottish vocational qualifications in the UK the focus is on

occupational competence, which is defined as ‘the ability to apply knowledge, understanding, practical and thinking skills to achieve effective performance to the standards required in employment. This includes solving problems and being sufficiently flexible to meet changing demands’ (QCA, 1998, p. 13). Based on a review of 100 NVQs and SVQs another brochure of the Qualifications and Curriculum Authority stresses that the standards are ‘expressed in outcomes [...] [which is] the required end result for the assessment of competence’ (QCA, 1999, p. 6).

The standing conference of State Ministers of Education and Culture in Germany (*Kultusministerkonferenz* – KMK) defines action competence (*Handlungskompetenz*) as the readiness and ability of a person to behave in vocational, societal and private situations in a professional, considered and responsible manner both towards him/herself and society. This competence concept is differentiated according to domain, personal, and social dimensions (KMK, 1996/2000).

In order to find out similarities and differences in these concepts they will be analysed according to the levels of the general diagnostic framework. The results are laid down in Table 1.

Table 1: Comparisons of concepts

Level	Internal conditions (knowledge, capacity, motives, etc.)	Actual episodes (action, behaviour, information)	External conditions (situation, task, requirement, outcome, product)
Definition			
European Commission (2001)	• capacity • experience • knowledge • qualifications	• use (effectively)	
Björnavold and Tissot (2000)	• capacity • know-how • skills • qualifications • knowledge	• use	• occupational situations and requirements (usual and changing)
France (Gilbert, 1998)	• knowledge (theoretical, applicable) • ability	• behaviour and motivation, structured (mastering)	• specific situation
NVQ-UK (QCA, 1998)	• ability • knowledge • skills (practical and thinking)	• apply understanding • problem solving • being flexible	• effective performance • demands (changing)
Action competence (KMK, 1996/2000)	• ability	• readiness to behave (professional, considered, responsible)	• situations (vocational, societal, private)
Source: Straka, G. A.			

Most of the definitions analysed span all the three levels but in varying differentiations; only the EU-definition does not include external conditions. Competence can be considered as a relational concept, bridging all the three levels involved in the individual-environment interaction. This has a strong impact for the public discourse: if the level and/or its elements are not specified, the audience may not be sure what the subject matter of a discussion is like. As a consequence, the probability of misunderstanding might be higher than of understanding. The result might be a pleasant and stimulating exchange of vague visions till others arise and dominate but no systematic empirically grounded insights and practical relevance.

In the context of measurement, behaviour is an indicator of action and the individual internal conditions. Bearing in mind that behaviour and action are transient, stimulated by the personally perceived environment and created by the internal conditions of the individual, there are two roads to specify competence:

(a) with situations, tasks or/and requirements (including the outcomes and/or criteria) for an occupation or an occupational function;

(b) with the knowledge, skills, abilities or more general the internal conditions necessary for an occupation or an occupational function.

In the first instance, much work has already been done in VET concerning occupational classifications (e.g. the international standard classification of occupations, ISCO, of the ILO, or VET curricula). However, it must be determined if they are adequate for measurement purposes. For the second approach, empirically tested notions, assumptions, or theories of development of internal conditions in occupational fields are required. Weinert (1999) argues that this context brings up a problem not yet solved in the 100-years history of scientific psychology. Therefore, he recommends the first approach which appears to be more functional (see NVQ). Eraut (2003) comes to the same solution but from a different perspective. He states that many definitions of competence can be differentiated, whether they are individually or socially situated. He prefers the second type, 'because public discourse on competence assumes that it is central to the relationship between professionals, their clients and the general public' (Eraut, 2003a, p. 2).

4. Procedures of assessing competence

According to Wolf (1995, 1998) four features are affiliated with competence-based measurement and assessment:

- (a) the emphasis on outcomes, specifically, multiple outcomes, each distinctive and separately considered;
- (b) the belief that these can and should be specified to the point where they are clear and transparent; that assessors, assessees and third parties should be able to understand what is being assessed, and what should be achieved;
- (c) the decoupling of assessment from particular institutions or learning programmes;
- (d) the idea of real-life performance essentially in non-academic fields.

These features raise the question of how to measure and evaluate competences. Direct and/or indirect observation seems to be the 'natural' and non-reactive approach. Direct observation relates to (visible) behaviour in specified work situations or generated products; counselling a client to choose a tour package or a menu in the tourism business, or realising a task on demand or a work sample are instances of this. Indirect observation refers to descriptions of activities, for example documented work samples including video tapes, project documentation and references. Such documentations may become a part of a portfolio, documenting increase in occupational competence. It is the purpose and the assortment that differentiates a portfolio from a collection or folder (Linn and Gronlund, 2000).

The observations, which are the basis for measurement, have to meet quality characteristics, some of them being accepted worldwide in both theory and practice of assessment (AERA et al., 1999 or DIN, 2002). The most important quality characteristics are validity, reliability, objectivity and, in addition, fairness, and usability (Linn and Gronlund, 2000; Lienert and Raatz, 1998; QCA, 1999; Schuler, 2000; O'Grady, 1991).

The criterion for validity is the degree to which the approach measures what is supposed to be measured (e.g. O'Grady, 1991). The recent and revised 'standards for educational and psychological testing' (AERA et al., 1999) define validity

as the degree to which accumulated evidence and theory support specific interpretations of scores. The bonding with theory is gradually reduced to evidence in specifying validity as the degree of the adequacy and appropriateness of the interpretations and conclusions drawn from observations or scores (Schuler, 2000).

Different aspects are added within the general notion of validity. Construct validity relates to the inference made about the perspective of adequate mapping of the individual internal conditions and their interrelations on the basis of the observations and scores. As an example, the assumption that information-creating actions consist of a dynamic interplay between learning and metacognitive control strategies, motivation and emotion (Figure 1) is to be tested with two steps. First, factor analyses should verify that each item of the measurement scale has an exclusive and important loading on the factor (= construct, e.g. motivation and not emotion). Second, structural analyses should at least reveal positive relations between these constructs (e.g. Straka and Schaefer, 2002). Referring to the general diagnostic framework or the assessment triangle (Section 2.1), the interpretation rule is on the test band.

The question whether the tasks represent the requirements of an occupation refers to content validity. Providing reasons why the documented practice project corresponds to the reference project in the German IT further education concept is an example of the implementation of such a quality criterion (Section 4.8). Prognostic validity refers to the accuracy of the prediction that can be made on the basis of the observations. Prognosis validity investigates, for example, the correlation between the methods used for hiring employees and their future career paths (Table 2 in Section 4.9).

Reliability expresses the degree to which scores for a group of test takers are consistent over repeated applications of a measurement procedure and hence are inferred to be dependable, and repeatable for an individual test taker (AERA et al., 1999). More generally, this quality criterion expresses the degree to which the same results are achieved irrespective of when, where or by whom a person is assessed (O'Grady, 1991). A

prerequisite of high reliability is that the subjective factors related to the assessor and unsystematic differences in administering and analysing are eliminated, minimised or held constant. These aspects are covered by objectivity, which is sometimes subsumed under reliability. A different quality criterion receiving more attention nowadays is fairness, which focuses on avoiding discrimination due to circumstances individuals cannot control. Usability addresses the practicability and, ultimately, the costs of measurement procedure.

For each of these criteria of measurement quality a sophisticated methodology for empirically based assessment has been developed since the Binet-Simon first IQ test in 1905. Furthermore, these criteria are interrelated, i.e. low objectivity influences reliability which itself influences validity (reliability is a necessary but not a sufficient condition for validity). In assessment practices in general, and also in VET, a reasonable balance between potential validity, reliability, costs and usability is essential. Fairness is becoming a major issue in scientific discourse and in some recent publications fairness is regarded as an essential part of a comprehensive view of validity (Linn and Gronlund, 2000). This should be borne in mind when evaluating a measurement procedure.

A selected sample of procedures for measuring and evaluation of competence used in the EU will be presented next with the aim of providing an impression of the range of approaches as well as their common features and differences. Afterwards, these procedures will be analysed and evaluated against the background of a general diagnostic framework (Figure 1) and the quality criteria for measurement described above.

4.1. The bilan de compétences (France)

The *bilan de compétences* is a broad diagnostic procedure taking place in accredited assessment centres. The bilan is a tool of feedback, flexibility and mobilisation of national continuing education policy. It has been a legal instrument since 1991 with several modifications since then (the latest was the decree for social modernisation in 2001). The main aims affiliated with the *bilan de compétences* are:

(a) strengthening self-responsibility for the occupational career of the individual;

(b) institutionalising the political goal of lifelong learning;

(c) enhancing the experiential knowledge, the transfer knowledge and the performance aspects in ones occupational biography (*acquis professionnels*).

This last aim targets workers with low formal education to establish a counterweight to the dominance of diplomas and formal certificates of qualifications job career paths in France (Thömmes, 2003).

The competence-based measurement of the *bilan de compétences* does not represent a uniform method or a distinct instrumentation. Different approaches are used, with common steps and phases. The starting point of the *bilan de compétences* is an occupational and/or personal desire to change, especially for people with low formal educational status. Knowledge, skills and other potentials acquired during working life are diagnosed, although formal qualifications and certificates are also considered.

The *bilans* are carried out in *Centres interinstitutionnels de bilans de compétences* (CIBC). Participation is voluntary and free of charge for the assessee. The result is given only to the participant, even though the employer contributes directly (time away from work) and indirectly (contributions to continuous training) to the procedure. The process of measurement consists of three phases, altogether taking 20 to 24 hours, with breaks between. On average about four to six weeks are necessary.

Phase 1 covers prearrangement of the contract (approximately four hours), fixing aim and areas, and compiling individually the repertoire of methods to be used.

Phase 2 is execution of the arrangement (12 to 16 hours): interviews to reconstruct the occupational biography; tests in areas such as intelligence, interests, learning ability, motivation, skillfulness and concentratedness; standardised personality inventories, simulations exercises, business games; and evaluation of the learning and developmental potentials.

Phase 3 is the final session (approximately four hours). This presents the results as a whole on the basis of a strengths and weaknesses profile and discusses the results with the intention to create a written outline for a project of an occupational change. There is also the arrangement of a follow up meeting about six month afterwards (Thömmes, 2003).

People conducting this procedure have to be

qualified (higher education degree in psychology, education, medicine and the like) and they are allowed only to use methods ensuring reliability and common scientific standards.

However, in reality, there is still a large discrepancy between educational policy aims and reality. The number of realised *bilan de compétences* decreased from 1993 (126 000) to 2000 (78 788) (Gutschow, 2003) and the proportion of unemployed persons using this instrument is above average. The future will show if the new law from 2001 initiated a change.

4.2. The national vocational qualifications (NVQ) (UK)

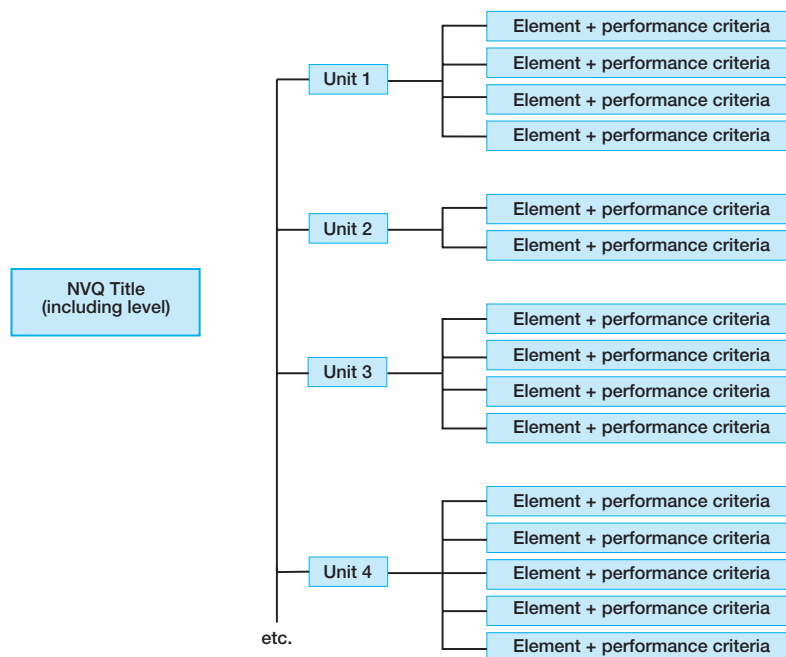
During the period 1980 to 1995 there was a strong cross-party consensus that the myriad of self-regulating awarding bodies should be reduced. The result was a competence-based system, the NVQ, in England and Wales. Employer-led organisations (lead bodies) define standards which embody and define competence in the relevant occupational context

based on a functional analysis of occupational roles (Wolf, 1995, p. 15 et seq.). Each NVQ covers a particular area of work, at a specific level of achievement related to the five levels of the NVQ framework. Levels 1 and 3 are presented as examples:

- (a) Level 1: competence which involves the application of knowledge in the performance of a range of varied work activities, most of which may be routine and predictable;
- (b) Level 3: competence which involves the application of knowledge in a broad range of varied work activities performed in a wide variety of contexts, most of which are complex and non-routine. There is considerable responsibility and autonomy, and supervisory competence in some capacity is often required.

Each level when applied to an occupation such as 'providing financial services (banks and building societies)', is divided into units, for example 'sell financial products and services' (CIB, p. 72) ⁽⁸⁾, for which standards have been developed. Each standard consists of an element and performance criteria, for example 'identify buying signals and correctly act on these' (CIB, p. 75, 3.2.5). The structure of an NVQ title is represented below.

Figure 2: The structure of a NVQ



Source: Wolf, 1995, p. 19

⁽⁸⁾ A unit itself may be allocated to different levels; for example unit 3 'providing financial products and services' of the NVQs 'Providing financial services' (banks and building societies) is allocated to level 2, 3 and 4. (CIB, n.d.). This unit is to be found in the Annex 3.

Extracts of the NVQ for 'providing financial services (banks and building societies)' are presented in the Annex. The standards for this sector comprise the levels 2, 3 and 4 of the NVQ system, with a total 64 units. The units are grouped in nine functional areas: sales and marketing, customer service, account services, mortgages and lending, financial advice, administration, asset management and protection, organisational effectiveness, and resource management.

The 33 units of Level 3 for the NVQ 'providing financial services' are subdivided into a mandatory set and two sets labelled as 'groups' from which units can be chosen according to regulations. Two units belong to the mandatory set (unit 45, contribute to a safe, secure and effective working environment; and unit 55, manage yourself), five units to group 1 (e.g. unit 15, maintain and improve customer service delivery) and 26 to group 2 (e.g. unit 2, sell products and service over the telephone). To gain the full NVQ at level 3, a total of eight units must be achieved: the two mandatory units, one unit selected from group 1 and the remaining five units from group 1 or 2.

As an example, unit 3, sell financial products and services (see Annex for the complete unit), relates to non-mortgage and non-regulated product and services, such as non-regulated insurance products investments and foreign currency. The assessee must show an ability to identify customer needs and promote suitable products and services by presenting them accurately and seeking customer commitment to buy (CIB, p. 72).

Unit 3 itself consists of three elements: (3.1) establish customer needs; (3.2) promote the features and benefits of the products and services; and (3.3) gain customer commitment and evaluate sales technique.

For each of these elements the assessee must demonstrate the following knowledge and understanding: 'which products and services you are authorised to promote, key features and benefits of products and services you are responsible for promoting, the types and form of buying signals your customers might use (three specimen out of nine statements)' (CIB, p. 72).

For element (3.2), promote the features and benefits of the products and services, nine performance criteria are specified, for example, '(3.2.1) explain clearly and accurately the features

and benefits of products and services that most closely match your customer's needs; (3.2.5) identify buying signals and correctly act on these; or (3.2.9) record all relevant information accurately' (three of nine performance criteria statements) (CIB, p. 76).

The NVQ assessment itself is based on two types of performance evidence: '(1) Products of the candidate's work, e.g. items that the candidate produced or worked on, or documents produced as part of a work activity. The evidence may be in the form of the product itself, or a record or photograph of the product. (2) Evidence of the way the candidate carried out activities or evidence of the processes in demonstrating competence. This often takes the form of witness testimonies, assessor observation, authenticated candidate reports of the activity, or audio/video recordings' (QCA, 1998, p. 19).

In unit 3 the following evidence requirements are listed: 'You must be able to demonstrate your competence at promoting the features and benefits of your products and services through: your sales performances with customers, and your corresponding customer records' (CIB, p. 75). The evidence criteria are further specified with appropriate examples listed: interview/discussion notes and records, referral documentation, where sales opportunities were passed to a relevant colleague, and personal customer records.

The 'element' also specifies the knowledge and understanding considered as necessary for performing according to the standard in this area. Examples for the descriptions are: 'Make sure you can answer the following questions, e.g.: what is the difference between "features" and "benefits"? How does this apply to the different products and services you deal with? What sort of buying signals might customers use to show their interest, or disinterest, in the products and services being offered? How could you deal with the different types of buying signals?' (CIB, p. 76).

These unit-based qualifications are open to everyone. There is no need to have a prior qualification to start an NVQ and because each is based on a number of units, it is possible for individuals to build up qualifications at a pace that suit them. Individuals can then achieve certificates for specific units or a full NVQ when all the units have been completed. There are no time

limits for completing NVQs. The NVQ assessment is generally carried out in the workplace, with the emphasis on outcomes rather than learning processes (CIB, p. 1).

Such a procedure meets, to a degree, the requirements for competence-based assessment:

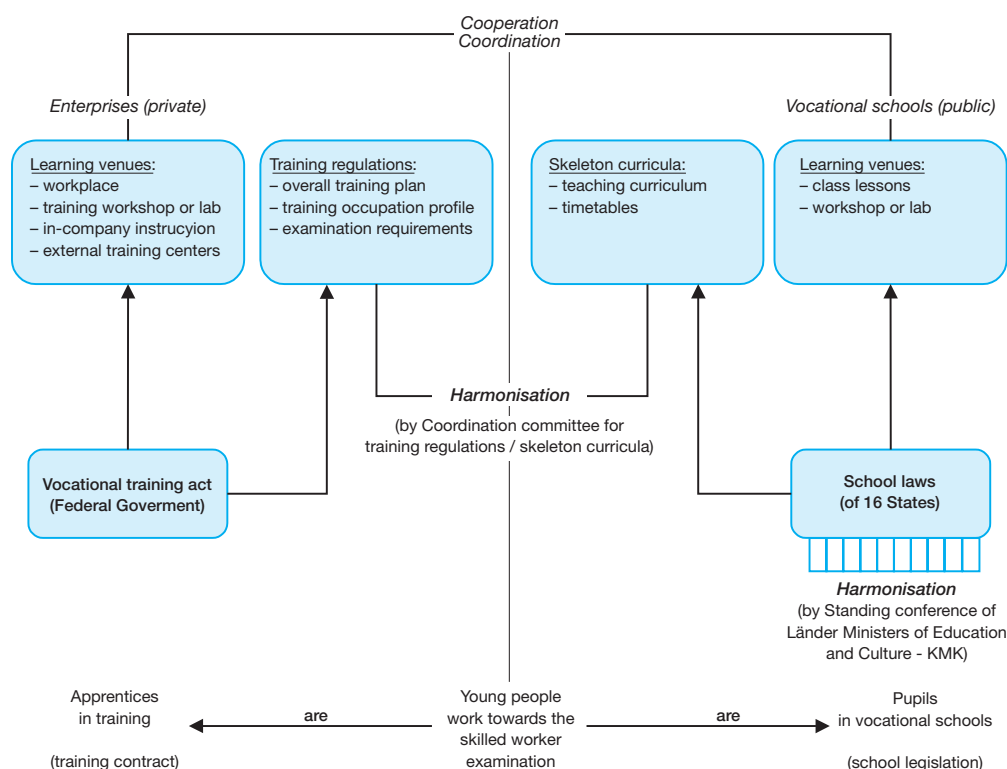
- (a) one-to-one correspondence with outcome-based standards;
- (b) individualised assessment;
- (c) competent/not yet competent judgements only;
- (d) assessment in the workplace;
- (e) no specified time for completion of assessment and no specified course of learning/study (Wolf, 1995, p. 20).

4.3. Assessing action competence in Germany

In 1996, Germany shifted to action competence, the development of which, however, is still bonded to explicit training in company and education in

school. The apprentice has to be simultaneously active in both venues for a fixed time, ranging from two to three and a half years. During this transition phase from school to work the young persons are company employees on the basis of a training contract and pupils in vocational schools (based on school legislation) at the same time. The goals for training and learning in these private enterprises are set in the training regulations consisting of an overall training plan, a training profile and examination requirements. The basis for the training regulations is determined in the 1969 Vocational Training Act of the Federal Government. Learning arrangements in the vocational schools are class lessons, workshops or labs. The teaching takes place according to state curriculum. Although the 16 German States enjoy full responsibility in education, two harmonisations are effected by the standing conference of State Ministers of Education and Culture (KMK). One is the coordination of the curricula of the 16 States; the other is tuning the training regulations to the skeleton curricula (*Rahmenlehrplan*) for vocational schools (Figure 3).

Figure 3: Overview of the structural features of the dual system



To demonstrate assessment subjects some extracts from training regulations and school syllabus, and from the profile for German bank clerks, follow.

The occupation profile is divided into the field of activity and 17 occupational skills. In the field of activity it is noted that bank clerks work in all the fields of business that lending institutions are engaged in. Their duties involve acquiring new customers, providing counselling and other services to existing customers and selling banking services, particularly standardised services and products. [...] The occupational skills of bank clerks is specified as follows: 'advise customers on selecting an appropriate type of account; [...] advise customers on investment possibilities in the form of shares, bonds, investment fund shares; [...] sell investment products; [...] evaluate the costs and revenues arising from business relationships with customers; [...] are competent in communicating, cooperating with others, solving problems and taking decisions' (Bundesanzeiger, 1998, p. 21).

The occupation profile, giving a condensed overview about the skills and knowledge of the occupation bank clerk, is made more concrete with the overall training plan (the training company is responsible for) and the skeleton school curriculum ⁽⁹⁾. According to the overall training plan for investment in shares, the following skills and knowledge should be acquired: '(a) inform the client about investments especially in shares, bonds and investment fund shares; [...] (c) assess the benefits and risks of investment in shares; [...] (f) answer client's questions concerning costs of buying and selling shares; [...] (i) describe derivatives and their risks in general' (Bundesanzeiger, 1998, p. 8).

Because of the shift from subject- to competence-orientation by the German vocational school system, the curricula should no longer consist of specifications of occupation related contents and skills but of learning fields (*Lernfelder*) ⁽¹⁰⁾ which are related occupational tasks and business processes. Each learning field is divided into specifications of objectives and contents.

For bank apprentices, the skeleton school curriculum consists of 12 learning fields for which some samples are given for learning field 4 (see Annex for a complete description): 'Offering long and short term investments in shares and bonds' comprises 100 of the 320 school hours of the first year of schooling. Examples of objectives are: The students identify client's signals for needs and motives for investment, communicate instruments of financing – client-oriented [...] use product-related calculations; explain services connected with investment decisions; describe risks associated with investment decision; heed regulations for protection of investors. Concerning contents, specifications like the following ones are specified: investment on accounts – like saving accounts [...] bonds, shares, investments funds shares, rights, risks, [...] basics and principles of investment consulting (Bundesanzeiger, 1998, p. 18).

In the German dual system of VET there is an interim and a final examination. Because the final examination comprises more assessment procedures than the interim one, only the final examination will be described. It consists of written tests in the following domains: business administration (180 minutes), control (90 minutes), and macroeconomics and civics (90 minutes). The tests consist of tasks related to practice (reality of interrelations) in selected contents to demonstrate analysis and understanding. Fixing them is the responsibility of federal and/or local examination boards in which the social partners are represented (employers, unions). For the bank clerk, the federal examination board recommends 60 % closed (multiple or other choices) and 40 % open tasks related to a national content catalogue (*Stoffkatalog*) developed and fixed by the same board, calibrated with the overall training plan and the skeleton school curriculum.

There is also an oral examination consisting of a simulation of client counselling (20 minutes and 15 minutes preparation; selected, by regulations fixed areas of content). As an example of the criteria and the way performance is measured and evaluated an extract is presented:

⁽⁹⁾ 'Skeleton' because of the sovereignty of the 16 States in education. Each State can develop its own syllabus along the skeleton school syllabus.

⁽¹⁰⁾ For a critique e.g. Straka (2001, 2002); Straka and Macke (2003).

Figure 4: Excerpt from an observation and evaluation sheet for client-counselling of bank clerks

Assessment Criteria (sample)	Remarks						
A. Action competence							
I. Interaction behaviour							
• produces interest in the client	_____						
• establishes contact with the client	_____						
• creates agreeable atmosphere	_____						
II. Informational and analytical behaviour							
• questions and analyses customers needs	_____						
• keeps the thread	_____						
•	_____						
III. Selling behaviour							
• persuades the client with arguments	_____						
• makes to the deal single minded	_____						
B. Subject area competence							
<table border="1"> <tbody> <tr> <td>A. Action competence</td> <td>_____</td> </tr> <tr> <td>B. Subject area competence</td> <td>_____</td> </tr> <tr> <td>Total points</td> <td>_____</td> </tr> </tbody> </table>		A. Action competence	_____	B. Subject area competence	_____	Total points	_____
A. Action competence	_____						
B. Subject area competence	_____						
Total points	_____						
Remarks:							
100–92=1	91–81=2 80–67=3 66–50=4 49–30=5 29–0=6						

Source: DHT, 1999

Such a simulation and evaluation of a client-clerk communication is a novel aspect of the shift of German VET to action competence in 1996 and the differentiation into domain, personal and social dimensions. A similar approach was implemented in the reform of commercial VET in Switzerland since 1999, where these three inter-related dimensions (domain, technique, and social) were integrated with the competence cube ⁽¹¹⁾ http://www.igkg.ch/deutsch/rkg_kompetenzwuerfel.htm

4.4. Social and methodical dimensions of competence (DaimlerChrysler)

Assessments are also developed and practised by companies. However, only a few have been published as these assessments are generally for internal use only. An available example was developed at DaimlerChrysler in Germany. It focuses on social and technical aspects of competence (Weisschuh, 2003).

⁽¹¹⁾ Available from Internet: http://www.igkg.ch/deutsch/rkg_kompetenzwuerfel.htm [cited 2.3.2004].

The assessment concept is named ‘training in dialogue’ (*Ausbildung im Dialog – AiD*) and, after a trial phase, was introduced company-wide around 1999. It is still in use and it is being considered whether to use this method for continuing assessments in both company education and training (Weisschuh, 2003).

‘Training in dialogue’ consists of an observation sheet to be completed by the persons in

charge of human resource development in order to record non-validated observations of apprentices’ behaviour (see element (3) in Figure 1), a sheet for apprentices to assess their own competences and a dialogue sheet to fix target agreements between trainer and trainee.

As an example of the assessment sheet, the items and ratings for the key competence ‘ability to cooperate’ are presented.

Figure 5: Rating scale for the key competence ‘ability to cooperate’

0	10	20	Ability to cooperate
30	40	50	
60	70		
☐		☐	
0	10	20	☐ Remarks/reason:
30	40	50	
60	70		
☐		☐	
0	10	20	Respects the opinion of others
30	40	50	
60	70		
☐		☐	
0	10	20	➤ Takes the opinion of others seriously ➤ Checks one’s own point of view within the conversation ➤ Stands by common decisions ➤ Treats others in an open and fair way
30	40	50	
60	70		
☐		☐	
0	10	20	Correct behaviour and support of others
30	40	50	
60	70		
☐		☐	
			➤ Establishes and supports contact with others ➤ Takes stock concerning others ➤ Refers to her/his own knowledge to that of others ➤ Support others ➤ Integrates outsiders into the trainee group

Source: Weisschuh, 2003

The trainer evaluates the trainee with respect to the different key competences between 0 (does not fulfil the requirements at all) and 70 (exceeds the requirements). In the same way, the apprentice rates her/his own competences. After that, both sides can come together to compare and discuss the ratings. In case of disagreement, the trainer can use only the behaviours written down by her/him on the behaviour sheet. The meeting should end with fixed target agreements recorded in the dialogue sheet.

4.5. Competence at work: a case in the Netherlands

In 1999, the advisory committee for education and labour market published a proposal under the name ‘shift to core competences’. The employers’ key argument is that people should be skilled in doing a job and that these skills should preferably be acquired outside the school world of disciplines and subject matter. This kind

of reasoning has led to detailed standards of occupational competence in each industry sector, forming the basis for vocational credits. Furthermore, the decoupling of learning process and assessment was one of the principles of the qualification system, laying the ground for a new kind of development: learning competence based on experiential learning.

A cooperative pilot project at Frico Cheese, within the Leonardo da Vinci programme, aimed to test a procedure of assessing prior learning and to validate competences relative to the qualification structure in agriculture vocational education. The motive for Frico Cheese to take part was based on upgrading the workforce. Employees had been working in the company for many years with virtually no additional training or education; many had bad experiences with full time school learning processes in the past. Accreditation of achieved competences (AAC) seemed to be a better way of reducing employee resistance to learning than formal processes of schooling or training. The human resource department of Frico Cheese decided that the accreditation of achieved competences procedure should be easily accessible to employees and carried out at the workplace by an internal assessor.

'A candidate is said to be competent when he or she is able to perform in realistic work situations according to the standards defined by the lead bodies in that specific work domain' (AAC-project, 2001, p. 41). The assessment process consists of two phases: evaluating a personal portfolio and demonstrating one's competences in an authentic work situation (= on the job).

A portfolio contains descriptions of relevant experiences and diplomas of the employee obtained in a formal or non-formal way. This description is to be compared to the national standard. Conclusions are drawn regarding content and level of competences. This is a kind of matching process between qualifications demanded and competences offered. The procedure allows for the easy identification of the elements already included in the portfolios as well as of those that have to be assessed by performance-based assessment tasks. Based on these outcomes, a decision is made whether the employee is allowed to continue the procedure.

Using a checklist of authentic tasks, the assessor compares the performance with the standards. In a successive interview the candidate is asked to reflect on the performed tasks and to respond to questions on transfer of working methods and solutions to similar situations. A positive evaluation of this process leads to recognition of the competences and a certificate closes this process.

Some 77 of the lower educated employees in the assessment procedure received an average of three certificates without visiting the community college. However, it is difficult to measure this success, not knowing how many people are not selected to take part in the procedure, and not knowing what the reliability scores are for the different assessors, both internally and externally (Nijhof, 2003).

A study done at Frico Cheese after one year of the Leonardo da Vinci project records no promotion of employees; there was no link between the number of certificates and salary increase. Some of the employees felt it easier to get a job outside the company based on the certificates, but these were exceptions (Nijhof, 2003).

Most participants felt, according to the interviews, that learning on the job has helped to improve job performance; this is followed by technical training. General education was considered the least important. Employees with accredited competences are perceived as good, or even best, at their job. Nijhof (2003) however, raises the question whether a classic performance test would not have given the same result.

4.6. The *realkompetanse* project (Norway)

The *realkompetanse* project of the Norwegian Ministry of Education and Research started in August 1999 and formally came to an end in July 2002. Its aim was 'to establish a system that gives adults the right to document their non-formal and informal learning [without having to undergo traditional forms of testing]' (VOX, 2002, p. 9). *Realkompetanse* includes 'all formal, non-formal and informal learning [...] Formal learning⁽¹²⁾ is normally acquired through

⁽¹²⁾ It should be pointed that 'formal learning' is a metaphor. Learning itself is exclusively a personal matter whereas 'formal', 'non-formal' or 'informal' vaguely indicate some characteristics of the environment (Straka, 2002a; Straka, 2003b).

organised programmes delivered via schools and other providers and is recognised by means of qualifications or parts of qualifications. Non-formal learning is acquired through organised programmes, but is not typically recognised by means of qualifications, nor does it lead to certification. Informal learning is acquired outside of organised programmes and is picked up through

daily activities relating to work, family or leisure' (VOX, 2002, p. 11).

The validation process combines dialogue-supported performance assessment and portfolio assessment, possibly in conjunction with tests (VOX, 2002). It covers the following areas:

- (a) documentation methods in work life and in civil society/third sector;

Figure 6: **Competence card from the workplace**

Specification of social and personal skills		Level
Cooperation and communication	Cooperates on a daily basis with management and other members of staff. Able to contribute to finding solutions. takes part in some important decision-making. Able to make other realise the importance of following up procedures and routines.	
Effort and quality of work	Reliable, keeps deadlines and appointments. Responsible and conscientious. Able to handle longer work intensive periods. Quality conscious.	
Customer service	Open and outgoing. Achieves contact easily with other people. Good knowledge about the needs of membership businesses. Experience and skills in active recruiting of new in-service businesses.	
Initiative – flexibility – creativity	Flexible. Likes to take on new tasks. Able to come up with proposals to new solutions to e.g. the Quality Control system, or arrangements for following up apprentices.	
Work related to restructuring – acquisition/use of new knowledge	Has taken part in trial runs of a new system for wages and personnel administration, and has made suggestions to the improvement of this.	

Specification of management skills in position		Level
Staff and labour management	Responsible for the daily follow-up of an apprentice in the company.	
Training and instruction	Responsible for the information to schools and classes. Takes actively part in the establishing of training programmes in local in-service businesses. Has taught classes in clerical courses for adult apprentices and interns.	
Level A = Carries out elementary tasks under supervision Level B = Works independently within own area of responsibility		Level C = May hold professional responsibility, may council and advise Level D = Has a very good insight in subject area of profession, may be in charge of development on own workplace.

Place:	Date:	Signature of employee:
Place:	Date:	Signature on behalf of business/company:

- (b) validation with respect to upper secondary education;
- (c) validation for admission to higher education.

The documentation of non-formal and informal learning in the workplace consists of two parts: a curriculum vitae and a skills certificate. The curriculum vitae is similar to that put forward by the European Commission and the forum on transparency of qualifications. It includes: personal data; work experience (employer, position, period, areas of responsibility); education; valid licences and public approved certificates; courses (names, period, completed year, important content); other skills and voluntary work (type of skill and activity, skills and closer description of responsibilities); and additional information. The skills certificate or the 'competence card from the workplace' describes what the employer is able to do as a part of his or her job. The categories are: personal data; main areas of work responsibility with a closer description of responsibilities; specifications of professional skills needed to carry out main responsibilities; personal capability; social and personal skills; management skills; and additional information. These specifications are graded at four levels, from level A (carries out elementary tasks under supervision) up to level D (has a very good insight in subject area of profession, may be in charge of development on own workplace). The certificate is signed by the employee and on behalf of the business/company.

Another example for the documentation of non-formal and formal learning is the 3CV developed by a number of organisations of the civil or third sector. This instrument contains an introduction (in which the methodology for completion is described), an example of a completed form, a form ready for completion, and the option of creating one's own reference.

4.7. Recreational activities (Finland)

The recreational activity study book of the Finnish Youth Academy is designed as a tool to make learning visible in settings outside the formal system. The target group is all young people over 13 years. Up to March 2002 there were over 55 000 study book owners in Finland and over 250 formal educational institutions acknowledge the value of the entries in the book. This concept

of valuing learning results in non-formal settings is supported by the Finnish Ministry of Education and Culture and the Finnish Forest Industries Federation. The major youth and sport NGOs behind the development include the centre for youth work of the Evangelical Lutheran Church of Finland, the Nature league, Guides and scouts of Finland, the Finnish Red Cross and the Swedish study centre in Finland.

The concept of the recreational activity study book differentiates nine settings or types of learning activities. The study book itself is divided into nine categories, according to the nature of the learning activity: regular participation in leisure activities; holding positions of trust and responsibility within NGOs; activities as a leader, trainer or coach; participation in a project; courses; international activities; workshop activities (apprenticeship); competitions; and other activities.

Examination of the categories shows that there are non-formal and formal environments for learning. The most formalised form for learning is the category 'courses' which means organised, and often hierarchical, educational programmes offered by various youth and sport NGOs and other training providers. The eight other categories fall more or less under the umbrella of informal sets, in which the learning-by-doing approach is often the method for acquiring competences and skills.

Adults responsible for the records are instructed to focus on the learning and development of the young person, i.e. not only to document what has been done but also how it has been done. Various aspects are recorded in the book: organisation responsible for the activity; description of the activity; young person's self evaluation of learning; date(s) and duration of the activity; performance (success and development); signature of the adult person responsible; contact information of the adult; and stamp of the responsible organisation (Savisaari, 2002).

The entries in the book are always written by an adult person (over 18 years of age) who is either responsible for or well aware of the particular activity. The learners fill in the 'self-assessment of the learning'. The idea is to focus more on what and how things have been learned rather than only what has been done. The person countersigning the entry adds his/her contact information, in case someone wants to check whether the young person actually has participated in the activity or not.

Figure 7: Example page of the recreational activity study book

Type of activity: Holding positions of trust and responsibility with in NGOs	
Organisation in which the activity took place	Time/dates of the activity ____/____/____ - ____/____/____
Position of the young person in the organisation	In average _____ hours per week/month Successes and competences acquired
Description of the activity	Place Date
Young persons self – assessment of the learning	Signature of the person responsible of activity
	Contact information of the undersigned person
	Position of the undersigned person

Source: Savisaari, 2003

4.8. New ways of assessing competences in information technology (IT) in initial and continuing VET in Germany

A new type of final examination was introduced in Germany in 1997 for recognised information technology occupations (*IT-Berufe*) within the dual system. Besides subject matter, macroeconomics and civics, two 'holistic tasks' (*ganzheitliche Aufgaben*) are to be solved in 90 minutes each (written practical part). The former practical part of the examination – for example, simulation of client counselling for bank clerks – was replaced by a

project to be realised in the company of the apprentice (written-oral part). This project work consists of an authentic assignment order or a defined partial task to be realised within 35 hours of work. The proposed project has to be accepted by the examination board and the documented result of the project work has to be presented orally in front of the local examination board, followed by a professional discussion.

A new IT continuing education system became active in May 2002 based on these recognised IT occupations of the German dual system (GDS) ⁽¹³⁾. The system consists of within-company career paths and three levels:

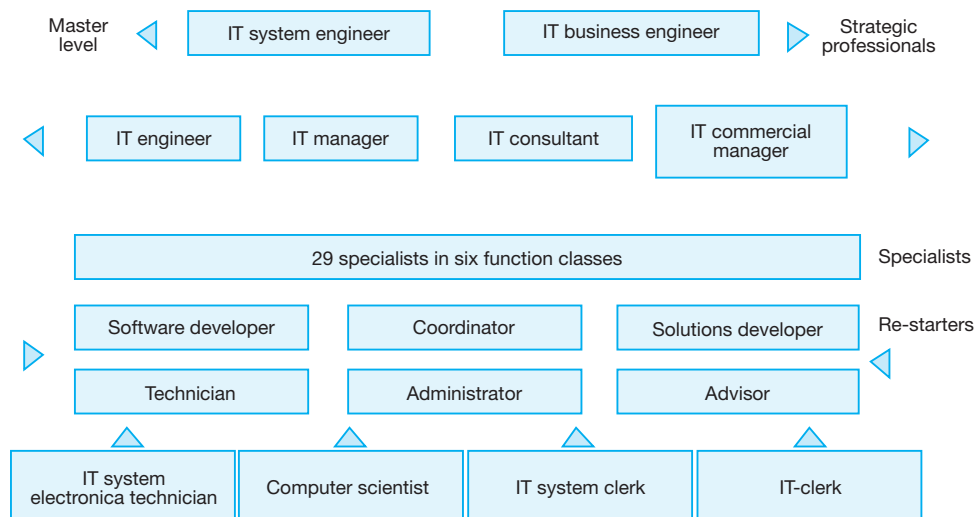
⁽¹³⁾ Available from Internet:
http://www.bmbf.de/pub/it-weiterbildung_mit_system.pdf,
<http://www.apo-it.de> and
<http://www.kib-net.de> [cited 2.3.2004].

- (a) the lowest level includes 28 special profiles, including software developer, network administrator, IT key account person and e-logistic developer;
- (b) the second level, with operative functions, is differentiated IT engineer, IT manager, IT consultant, and IT commercial;

- (c) the third level, with strategic functions, is divided into IT systems engineer and IT business engineer.

A certificate for each of these levels will be treated as equivalent to a BA or MA certificate in the higher education system via the European credit transfer system (ECTS) (Figure 8).

Figure 8: **Structure of the IT-further education system**

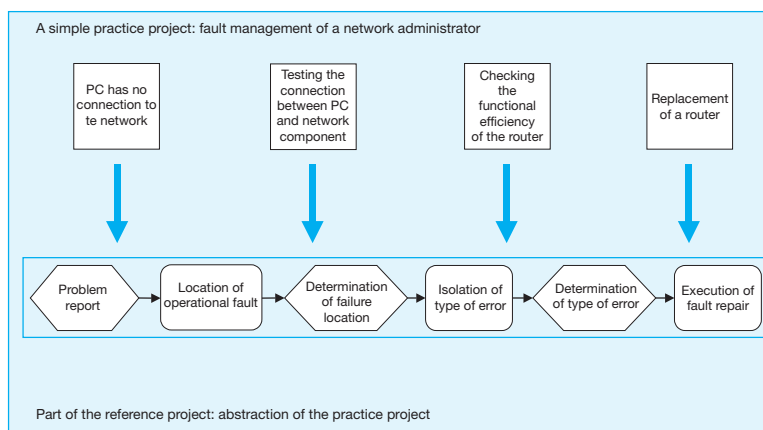


Source: Grunwald and Rohs, 2003

The basis for the certificates acquired in this system is a ‘work process oriented continuous training’ during which the employee works on actual requirements of company projects in order

to combine work and learning. The project work has to be documented, discussed with the coach in the company and checked against ‘reference projects’ (Grunwald, 2002, p. 179).

Figure 9: **From a practice to a reference project – an example for a network administrator**



Source: Grunwald, 2002, p. 170

Certification is run with a private certification body according to the general criteria for institutions certifying personnel according to the German Industry Norm (DIN EN 45013).

4.9. Analysis and evaluation

A review of the introduced sample of approaches to assessing occupational competences indicates a wide span of procedures. They range from broad, such as the *bilan de compétences*, to narrow targets, such as the rating scales for specific social and methodological dimension competence at DaimlerChrysler. In the general diagnostic framework, the foci of the approaches tend to be different. NVQs concentrate on the external conditions (i.e. tasks, solutions) to be solved and observable behaviour whereas, according to the German view of competence, assessment, actions and conclusion, internal conditions are the subjects of consideration.

All the procedures described tend to get away from the classical measurement, characterised by well defined internal conditions and tight tasks mostly formatted with multiple choice items. Instead they favour authentic and complex tasks from the shop floor. Combined with the constructivist approach and interconnected with the qualitative empirical research approach, introspection (self-assessment), peer assessment or observing on-the-job performance seems to be the first choice for measuring and evaluating competences, especially for softer dimensions like the social and communicative. But, no matter which approach to competence assessment is chosen, it should attain the highest possible degree of validity, reliability, objectivity, fairness and usability, i.e. the quality criteria of any assessment.

Introspection and self-evaluation are explicitly or implicitly biased, a phenomenon extensively investigated in survey research under the term 'socially desired answers' (e.g. Kerlinger, 1973). Another source of errors stems from features of general and job-related knowledge and skills. Personal explicit knowledge as a part of the internal conditions may become a matter of routine especially if practised over a long period. The opposite takes place when people learn to act successfully but are not aware of the knowl-

edge base of their actions. In both cases an identification of the knowledge incorporated in such activities makes self-evaluation impossible. Both phenomena are linked with the concepts of implicit and tacit knowledge, the latter characterised by 'we know but cannot tell' (Polanyi, 1966). These features of actions and knowledge set narrow borders to introspection as a basis for assessment of vocational aptitudes.

Work documentation and portfolios are also given high attention in the context of competence-based measurement: see *bilan de compétences*, the case from the Netherlands, or the assessment approach of 2002 German IT-continuing education. Such products are not fugacious and they can be observed, analysed and evaluated independently by different assessors. However, the selection of work pieces to be documented may be biased, and there is no guarantee whether the assessee produced the outcomes on the basis of her/his competence and without help of others.

Portfolio procedures are not easy to handle because of the heterogeneity of information being recorded, and the variation of experiences of employees. Last but not least, there is little systematic or empirical investigation of whether portfolios are acceptable predictors of future performance, especially in the case of continuous changes in the work environment (Nijhof, 2003).

At a first glance, observation by external evaluators might alleviate the weaknesses of introspection and self-compilation of documented work samples. However, this type of observation carries its own bias, because people tend to reconstruct the surrounding world under the perspectives of their prior knowledge and cultural perspectives (e.g. Shephard, 2001; Bransford et al., 2000).

Observations in natural settings by others – peers, supervisors – are more likely to be unsystematic than systematic. This fact is taken into consideration in practice, by using specifications for recording and evaluating observations, such as the check list for assessing the simulated client counselling in the German final examinations for bank clerks (Section 4.3), the rating scales of 'training in dialogue' at DaimlerChrysler (Section 4.4), the criteria of the Norwegian competence card from the workplace (Section 4.6.), or the in the Finnish recreational activity study book (Section 4.7).

Basing external observation and evaluation on defined aspects or criteria is a way of handling such sources for errors. However, the core problem is still left: the processes of recording, summarising and evaluating take place in the brain of the assessors and, therefore, they cannot be cross-checked by outsiders to converge to objectivity. Beck (1987) lists 20 sources for errors aligned with rating and he warns against 'fantastic performances of observers'.

These methodological and theoretical considerations raise the question whether any empirical evidence exists concerning the measurement quality of assessments in VET. Unfortunately such evidence is scarce.

The *bilan de compétences* in France is neither dedicated to measuring potential nor to supporting decisions concerning personal or professional development. The proportion of subjective measurement is high and the introduced quality criteria are generally seen as not suitable (Thömmes, 2003). A similar view might have guided the Dutch Frisco Cheese project funded in the Leonardo da Vinci programme when Nijhof (2003) criticises the absence of considerations concerning reliability and validity.

Eraut points out that the demonstration of competence in company A does not guarantee that this competence will work in company B. According to his investigations, Eraut (1994, 2000, 2003b) competences are context- and situation-bounded and, therefore, transfer from A to B needs an additional learning process. The NVQs prioritise on-the-job learning and the term competence was selected for 'its declared purpose of accrediting only effective performance in the workplace; but NVQs also had to be national qualifications' (Eraut, 2001, p. 94). Therefore, a generic language is used to describe activities which might have the same function but are very different in character. 'This enabled formal comparisons between "equivalent" jobs in different contexts, but did not guarantee that competence in one context could be readily transferable to another' (Eraut, 2001, p. 95). Finally, an evaluation of high level NVQs indicates

the dangers of a fragmented approach to performance, and the limited development of expertise in assessment and learning support (Eraut et al., 2001).

In Germany there is considerable critical discussion about the pros and cons of assessing competence and quite a lot of changes in methods of assessing competences in different occupations have taken place over the last decade. However, the discussion, the introduction and the practice are predominantly based on speculative reasoning instead of systematic empirical evidence. Actual assessment practice in German VET seems to have more in common with the rituals for transition from apprenticeship (novice) to craftsmanship (expert) in the Middle Ages than with the potentiality of measurement approaches in recent times (Straka, 2003a; see discussion in Chapter 5).

A meta analysis of empirical findings in occupation-related proficiency was carried out to determine empirically based devices for the potential measurement quality of assessment procedures. It researched the prognostic validity (*Berufseignungsforschung*)⁽¹⁴⁾ of different selection procedures compared to future occupational success.

The correlation (r) between data about 'occupational proficiency' and later 'occupational success'⁽¹⁵⁾ – the indicator for prognostic validity – can range between '1' and '0'. '0' represents no and '1' perfect prognostic validity. Table 2 indicates that the range of prognostic validities of the presented approaches is quite large. The lowest prognosis factor for occupational success is chronological age (.01) and the highest are standardised cognitive skill tests (.51), structured interviews (.51), and work samples (.54). However, the general validity of work samples should not be overestimated because they are often very closely related to the requirement of a specific workplace. Therefore, it is not certain that the same knowledge and skills are transferable to another context, especially if they are tacit (Eraut, 2001; Straka, 2001b).

⁽¹⁴⁾ In this context it should be pointed out that a German standard for requirements, procedures and their application in job-related proficiency assessment was published in June 2002 (DIN, 2002).

⁽¹⁵⁾ It should be noted that 'occupational success' in the table above is predominantly based on evaluations by supervisors, a method whose measurement quality is questionable.

Table 2: **Validity of customary approaches**

Validity of customary approaches to diagnosing occupational proficiency and occupational success	
Approaches to diagnosing occupational proficiency	validity (r)
conventional interview	.14
personality analysis (in general)	.15
conscientiousness	.31
integrity test	.41
school grades	.30
references	.26
work sample	.54
biographic questionnaire	.37
assessment centre	.35
structured interview	.51
time of probation	.44
cognition skills tests	.51
graphology	.02
age	.01
work experience	.18
Source: Moser, 2003	

At first glance, a correlation of 0.5 within the boundaries of '0' and '1' seems to be quite reasonable. But a more powerful indicator is the common factor variance expressing the overlap or the communality of the variances of two variables (in our case 'occupational proficiency' and 'occupational success') (Kerlinger, 1973). It is calculated by squaring the correlation coefficient ($= r$) and in the case of $r = .50$ the result is 0.25, which expresses that 25 % of the common variance, for instance, 'occupational proficiency' and 'occupational success' are due to their commonness, and 75 % of the 'residual' variance might be ascribed to other factors such as motivation, workplace design, organisation structure, work climate. Even so, hiring people on the basis of a structured interview or a cognitive skill test is, on average, significantly better for occupational success than selecting them by age or graphology (Table 2).

The EU approach to assessment procedures indicates a certain preference for performance-based

assessment. The main differences between performance-based assessment and classical approaches are presented in Table 3.

Expectations for the performance-based assessment are extremely high but a reasonable decision for the one or the other is ultimately a matter of empirical proof. However, the problem underlined by Baker et al. (1993, p. 1210) remains: 'Although interest in performance-based-assessment is high, our knowledge about its quality is low'. Some considerations indicate that this approach does not only have advantages. Complex problems linked with significant degrees of choice may increase the number of steps required to solve a problem as well as the number of solutions. For example, the cognition technology group at Vanderbilt (CTGV, 1997) constructed 'authentic' problems with at least 14 steps required to solve them. From the angle of pure combinatorics there are $14!$ ($14 \cdot 13 \cdot \dots \cdot 1$) sequences or solution paths (permutations) possible. For content reasons most of them will give no sense but there may be a number of meaningful ways and solutions. Considering that there is not a single right or wrong solution, several parties rating independently are necessary to achieving objectivity and fairness. This approach – used in VET too – has consequence on usability and some biases and errors of measurement remain.

In parallel, multiple-choice items get much attention in public debates on assessment. Their validity has often been questioned compared with open ended items, especially in the context of performance and competence assessment. A methodological analysis in TIMSS-III-Germany of items measuring mathematic-science literacy found out that using multiple choice or extensive answer formats impacts on measured performance, but that it is not the case when multiple choice is compared to short answer format. However, independent of the format, the TIMSS-III items were a much better indicator of general ability in mathematics and science literacy. Therefore, the TIMSS analyses used an overall score ignoring the item format (Klieme et al., 2000). A recent study (Straka and Lenz, 2003) of commercial VET shows that students used more learning strategies for solving multiple choice items in the test for economic literacy (TEL), (Beck and Krumm, 1998) than for open ended teacher-made

items. Considering such evidence, it may be concluded that open-ended tasks are per se not more valid than multiple choice items.

To summarise, much supports the assumption that measurement quality is related to its format but possibly more to the content of the item. Consequently, there is a demand from Baker et al. (1993) for better research to evaluate the

degree to which newly developed or practised assessments fulfil the expectations. Similarly, Brown (1999) notes that multiple choice items ‘are much more sophisticated than many of us once believed (see the UK Open University for examples of best practice) and don’t always have to be self-generated by tutors’.

Table 3: Performance-based assessment versus ‘classical’ approaches

Features of tasks/situations	Classical approach	Performance approach
1. Task format	Closed (multiple choice)	Open ended
2. Required skills	Narrow, specific	High order, complex
3. Environment relation	Context free	Context sensitive
4. Task/requirement	Limited scope, single and isolated skill, short time processing	Complex problems, requiring several types of performances and significant time
5. Social relations	Individual	Individual or group performance
6. Choices	Restricted	Significant degrees
Source: Baker et al., 1993		

5. When should competence be assessed and by whom?

The approaches to measuring competence presented in Chapter 4 are embedded in different external conditions in which competence is developed and assessed. These range from the development of work-related capacity as a result of acting in a workplace during working life (e.g. NVQs, *bilan de compétences*, or the competence card from the workplace) to personal development initiated and supported by a combination of educational and authentic work settings in dual systems (e.g. the dual VET in Germany), or exclusive educational arrangements with real and/or simulated practice (e.g. technical or vocational schools).

Linked with such approaches are regulations for assessment and its timing. German dual VET, for example, has two fixed dates for assessment – in the middle and at the end of apprenticeship – whereas a *bilan de compétences* or an assessment in the framework of NVQs can be requested and run at any time.

The question of the scope of assessment has to be tackled in this context. In VET courses, the contents and the objectives of the whole course might be the subject of assessment, as in the dual system in Germany. In a training system organised with modules and credits, certificates or grades confirming the mastery of a set of defined modules can be combined according to regulations for certification. For example, to get a level three certificate for providing financial services, the mastery of a defined number of units ⁽¹⁶⁾ out of different sets (Section 4.2) has to be certified by an external assessor.

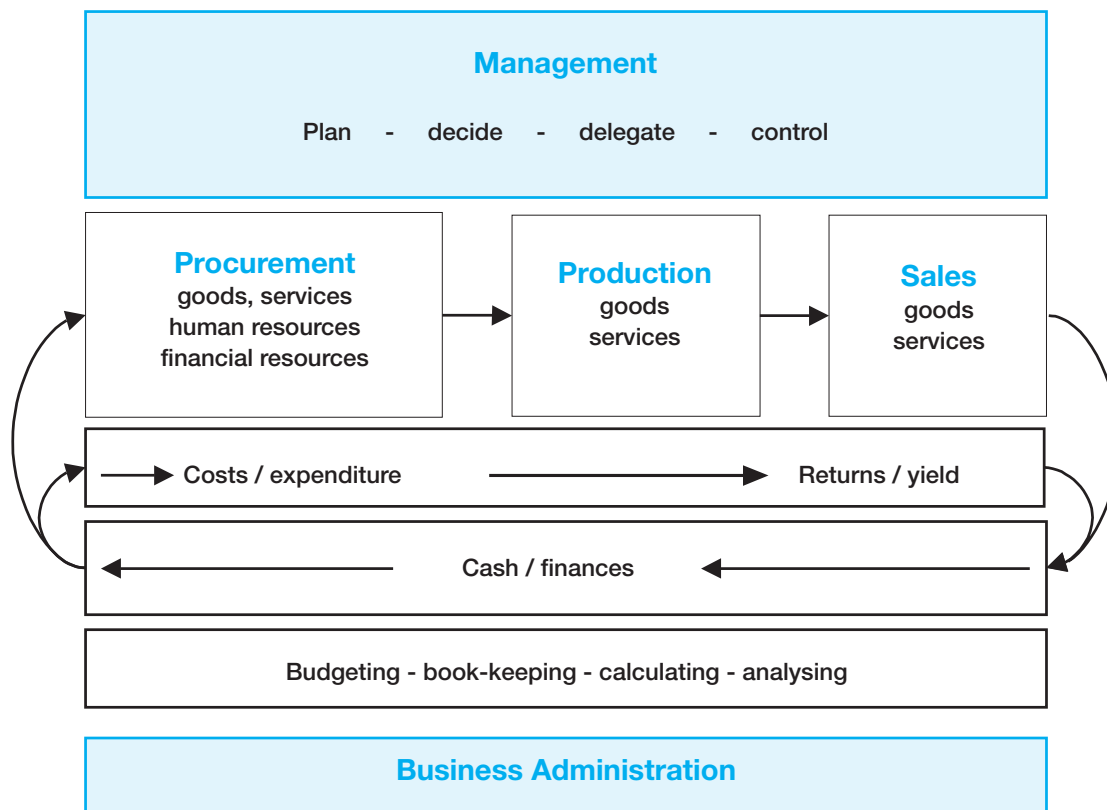
There are disputes about modularisation and assessment in the EU (Ertl, 2002). Discussion between supporters of modularisation and proponents of courses concentrates, above all, on fragmentation versus wholeness, which need to be balanced for business administration (Figure 1).

Enterprises are open, dynamic, complex, autonomous, market-oriented productive social systems (Ulrich, 1970) embedded in socioculturally shaped environments. This dynamic and complexity contrasts with the general model of the economic process (Porter, 1985; Wöhe, 2002) where production and selling of goods and services results from combining material and immaterial factors according to economic principle. In order to realise such goods and services, goods (raw materials, equipment), services and human and financial resources have to be procured. Every input and sale of goods has direct or indirect financial impacts with associated costs/expenditures, returns/yield and cash/finances. The whole process is initiated, planned, decided, delegated and controlled by management. The information needed for this function is provided by budgeting, calculating, accounting and analysing the flow of goods, services, costs, expenses, and returns (Thommen, 1991).

Regarding a module as a ‘subset of a learning program’ and a unit as ‘a coherent set of learning outcomes’ (Ertl, 2002, p. 143), and noting that there are no specified criteria for coherence and subsets, there is significant freedom in structuring teaching programmes and specifying and assessing learning outcomes. Elements such as management, procurement, or sub-elements of them such as planning, deciding or further subsets, might be the subject of units or modules which themselves are considered interrelated. Their objectives might be assessment and grading. Teaching and realising the objectives, however, may take various paths through these sets: planning, to controlling, to delegating and deciding, and vice versa; sales to production to procurement or vice versa; or planning of production, sales, etc. The structural organisation of

⁽¹⁶⁾ An interpretation of a unit as a module is possible because ‘it is unclear in many cases whether the words “unit” and “module” are used interchangeably or whether they mark a difference’. The situation is similar when organising the teaching process in modules is the issue (Ertl, 2002, p. 143).

Figure 10: General model of economic processes, a basis of business administration



these modules is different and the structure of the business process is in part hierarchical and inter-related. However, the processing itself is always sequential and linear, i.e. one item after another (as indicated above) which has some consequences for assessment too.

Figure 1 showed that the basis of measurement and evaluation are observations of products of actions and/or the behavioural part of actions. On this level the observations are similar, independent of whether the education and/or training is structured in modules or courses. For example, if bookkeeping or sales is on the agenda, the action episodes are the same. They even have a considerable overlap across different economic cultures, as was seen in a pilot analysis of the operationalised educational objectives for bank clerks in the German VET regulations and of the performance and evidence criteria in the frame of the NVQs for 'providing financial services (banks and building societies)' (Straka, 2002b; see Annex). Therefore, in both systems – German VET

and NVQs – similar observations might be the basis for assessment.

However, there are differences in the inferences drawn from observations. If the focus is purely on performance, the observation is sufficient if the task is solved or the process mastered by a person. But if the knowledge base for these solutions or of the processes is targeted, assumptions about the knowledge and its structure are necessary and have to be validated (= construct validity). For example, a purely performance-oriented assessor would mark her/his observation 'explain clearly and accurately the features and benefits of products and services that most closely match your customer's needs' (see performance criteria 3.2.1 in Annex) as a satisfactory, whereas an assessor focusing on internal conditions (knowledge, skills, etc.) might interpret the same observation as an indicator of declarative and procedural knowledge. As a consequence, an investigation of the effects of holistic or fragmented, modularised or

programmed, dynamic or static assessment is not a question of assessment but of the aims – or more generally the theory (Popper, 1989) – and the organisation of vocational training and/or education.

In this context, the question may be asked whether the best practice and exchange of experience models (*Modellversuche* in Germany) recommended between Member States (often in the case VET) on assessment procedures is a good alternative to systematic empirical research in this field? Generally, reports of best practice describe what was processed, the results and the conditions. In such documentation, relations between input, output and processes are considered in most cases on the basis of qualitative data, and according to specific interpretations of the constructivist paradigm; generalisations are not intended. Therefore, it is not surprising that transfer of best practice normally does not work. The situation and context tie of the practice make the transfer its own experiment (Eraut, 2000). On the other hand, from experimental and quasi-experimental research methodology factors interfering with the validity making the transfer impossible are known (Campbell and Stanley, 1963; Straka, 1974). Consequently, systematic empirical approaches with accurate research designs are to be preferred.

The question of when assessment should be organised is connected with the question of by whom the assessment should be done. As discussed in Section 4.9., self-assessment is the weakest approach for assessment. From the perspective of validity, preference should be given to assessment by others. Where tasks have a wide range of solutions, e.g. in the case of dialectical problems – neither the target nor the ways to reach it are defined (Dörner, 1976) – or approaches, several evaluators are usually engaged.

Involving social partners in assessment is not only a way to balance it more but also contributes to fairness. In the German dual system, for example, the participation of the chambers of commerce, the trade unions and the school as examiners is regulated by law. However, from a measurement angle the question is whether the social partners and their representatives guarantee more valid assessment than trained assessors.

Because there is no empirical data available to answer this question, a fictional reflection may

reveal some problems. Certificates and/or grades should guarantee that the assessed person possesses the skills, knowledge, and attitudes which make her/him eligible for an occupation or able to perform it successfully. If social partners are involved in defining assessment standards, they might do so representing divergent class positions, i.e. the traditional antagonistic ones of capital and labour. Their judgement might be influenced primarily by power relations and only after by questions of the validity, reliability and objectivity of such a standard. As a consequence, the certificate or grade defined might rather reflect the following metaphor: a dog, having its snout in the fire and its tail frozen, feels, on average, agreeably warm (Straka, 2003a).

In VET, peer or colleague assessment – especially in the case of group work – might be an alternative form of assessment by others. Such persons know each other and are familiar with work requirements and conditions. In addition, evaluations might be done on the basis of much more observations over a far longer time than punctual or even continuous external examinations. However, this ‘natural’ procedure bears some analogue risks as a retrospective field study found out. In the context of socialist production in the former German Democratic Republic the labour units (*Produktionseinheiten*) at the Rostock Neptun shipyard gained increasing autonomy for fixing work norms. An effect of this was that, during the 1950s, the over-fulfilment of the norms exploded dramatically; exceeding norms by 200 % or 300 % was not unusual. But, in reality, shipyard productivity decreased; this was a trend which may have contributed to the implosion of the former German Democratic Republic (Alheit, 1998).

In this context, one might wonder if external and punctual final assessments might not offer a better guarantee of validity? An answer might be derived from the German situation. Very often – especially when a new group of apprentices is hired – complaints about decreasing literacy in reading, writing, and arithmetic, etc., are heard from companies or their associations. The complaint may have some truth, as TIMSS and PISA indicate for certain segments. However, success rates in the final examinations are surprising. Regularly, they reach around 85 % of those who sit for final examination in the dual

system in Germany ⁽¹⁷⁾ (*Berufsbildungsbericht*, 2003, p. 99). Different interpretations of this are possible; starting a totally new career path motivates the young persons to close the gaps from former schooling, or the efforts in instruction and training in vocational schools and companies have remedial effects on the capabilities of apprentices. But another interpretation might be possible as well. The assessment procedures carried out by the social partners have more in common with the rituals of guilds of long ago than with modern procedures of measurement and evaluation (Straka, 2003a). This justifies advocacies for a VET-PISA (Pütz, 2002), and the related feasibility studies which are planned (BIBB-Workshop PISA-B 2003).

A VET-PISA might be in accord with the recent educational reforms, emphasising accountability with the following significant new features:

- (a) ambitious, world-class standards;
- (b) forms of assessment that require students to perform more substantial tasks (e.g. construct extended essay responses rather than select answers from multiple-choice items);
- (c) imposing high-stake accountability mechanism for schools, teachers, and sometimes students;
- (d) including all students (Linn and Gronlund, 2000, p. 5 et seq.).

However, such a kind of permanent VET-PISA would have dangers as well as advantages. According to the American education research association the following conditions have to be met in high-stakes testing:

- (a) protection against high-stakes decisions based on a single test;
- (b) adequate resources and opportunity to learn;
- (c) validation for each separate intended use;
- (d) full disclosure of likely negative consequences of high-stakes testing programmes;
- (e) alignment between the test and the curriculum;
- (f) validating passing scores and achievement levels;
- (g) opportunities for meaningful remediation for examinees who fail high-stakes tests;
- (h) appropriate attention to language differences among examinees;
- (i) appropriate attention to students with disabilities;
- (j) careful adherence to explicit rules for determining which students are to be tested, sufficient reliability for each intended use;
- (k) continuous evaluation of intended and unintended effects of high-stakes testing (AERA, 2000).

⁽¹⁷⁾ The range of success rates in 2001 included crafts at 80.6 %, industry and trade at 88.6 %, public service at 91.1 %, and overall at 86.1 % (*Berufsbildungsbericht*, 2002, p. 99).

6. Summary and conclusion

The emphasis on measurement and evaluation of competence is rooted in:

- (a) a shift from the predominant input to an output view of education, linked to accountability of educational systems and institutions;
- (b) a shift from subject to literacy orientation in general and vocational education;
- (c) a rise in the cognitive-constructivist paradigm with its orientation toward qualitative assessment methods as opposed to the behavioural-cognitive paradigm of traditional quantitative procedures;
- (d) greater recognition of skills acquired in non-formal and informal settings during life-time, linked with the request for innovative forms of certification.

Considering that any human performance is based largely on the knowledge, skills and motives of the individual, a general diagnostic framework was introduced. It differentiates three levels: internal (e.g. knowledge, skills, motives) and external conditions (e.g. situation, task, product) both bridged with the actual individual operations (e.g. behaviour, action). In this model, observing, measuring, and evaluating changes in the external conditions and the behaviour is a minor problem compared with that of making inferences on the non-visible part of the actions and the internal conditions. In order to solve this problem, explicit theories with well-defined constructs about domain specific features and the development of internal conditions, cognitive actions and interpretation rules linking observations with these constructs are necessary (Table 1).

Evaluation may have as a benchmark the measured characteristics of other persons (norm-referenced) or standards defined independently from measured characteristics of other persons (criterion-referenced). The latter is an appropriated, recommended, and used criterion in the context of VET.

Analysis of a sample of competence definitions used in the EU with the diagnostic framework revealed that the concept of competence is a

relational, one bridging all the three levels of Table 1 in most cases. Such a broad notion requires not only an accurate definition of the elements of the levels and their interrelations but also their systematic and empirical validation. Otherwise there is a significant risk of mutual misunderstanding in public and scientific discussions about VET competence, because it might often be unclear which level or element is under consideration. The result might be an interesting, even stimulating exchange of broad visions but of negligible scientific and practical relevance. Such a holistic view of competence might also become counterproductive to transparency and mobility, the officially stated objectives of the European Commission.

Practised assessment procedures in the EU were analysed against the background of the general diagnostic framework. The examples chosen included the *bilan de compétences* (France), the NVQ (England and Wales), different dimensions of action competence in the German dual system, assessing competences at work (the Netherlands), realkompetanse (Norway), the assessment of recreational activities (Finland), and valuing competences in the continuing IT-training (Germany).

The results of the analysis include an overwhelming orientation towards measuring and evaluation of performance in 'authentic' or 'natural' settings with standards of differing sophistication (e.g. NVQs) or prototypical project solutions (e.g. German IT-continuous education). There is also a preference for open-ended rather than multiple choice tasks, for complex rather than narrow and specific skills, for context-sensitive rather than context-free strategies, for complex problems requiring several types of performance and significant time rather than tasks with limited scope for a single and isolated skill to be solved in short time, and for significant degrees of choice rather than restricted choices.

Task format can affect performance but there is empirical evidence that the format is much less important than the task requirements in terms of content. Systematic empirical research on perfor-

mance measurement with 'authentic' tasks is required to examine whether they support the implicit and explicit expectations placed on them.

No published empirical evidence for the quality criteria of measurement, such as validity, reliability, objectivity, fairness and usability, can be found. However, summarised results from occupation-related proficiency research indicate a wide range of prognostic validity – from nearly 0 (chronological age) to .54 (work sample) – related to different approaches to diagnosing occupational proficiency (Table 2):

Different types of observation are used, ranging from introspection to observation, both direct (e.g. self-reports, participating observation) or indirect (portfolios). Since the act of recording, summarising and evaluating cannot be checked by outsiders, such approaches are subject to many potential errors of measurement, even when operational observation criteria are available. Recent research shows also that this type of assessment might carry sociopsychological constraints, especially if done by superiors (Krumm and Weiß, 2000). Therefore, the use of externalised work products or audio-visually recorded behaviour is recommended.

However, in the context of VET the ultimate goal is the sustainable change of personal characteristics. Hence these systematic and quantified observations become indicators of internal conditions. An approach such as the one chosen in the PISA study, that assesses the ability to complete tasks in real life, depending on a broad understanding of key concepts, rather than assessing the possession of specific knowledge in specific situations, is to be recommended. The domains are defined in terms of 'the content or structure of knowledge that students need to acquire in each domain [...] [see internal conditions]; the processes that need to be performed [...] [see actual episodes]; and the context in which knowledge and skills are applied [...] [see external conditions]' (OECD, 2001, p. 19).

This review covers only the tip of the iceberg for procedures, concepts of measurement of competence. The results underline the need for a review of VET competence assessment procedures from the methodological angle, confirming the conclusions of Björnavold (2000) in making learning visible in the field of learning in non-formal and informal settings. Starting with a

detailed description supported by concrete examples, the focus of further reviews of VET competence assessment should be on empirical evidence with respect to measurement quality criteria: objectivity, reliability, validity, fairness, and usability. If sufficient research findings are not available, empirical investigations concerning the quality of measurement of these approaches should be initiated.

This methodological undertaking has to be complemented by a conceptual domain-specific approach taking into account the non-visible parts of the actions and internal conditions and their development. Beyond VET, many such concepts from neighbouring disciplines are available (e.g. Weinert, 2001; Straka, in preparation). Their adequateness for VET should be examined instead of producing 'new' ones, in most cases not empirically validated. Such investigations could become a basis for a VET PISA in selected important forward-looking occupations or sectors under the auspices of the EU. It is time to stop initiating projects discovering the sociocultural idiosyncrasies in European VET systems and increase those mapping European VET on the basis of sound empirical data.

Valuing competences acquired in non-formal and informal settings also has elements of paradox. A rationale for this policy is that schools and their certificates have a marginal prognostic value for mastering life outside the formal settings (Resnick, 1987; Straka, 2003a). Another is that learning in out-of-school settings creates less inert knowledge than in classrooms (Cseh et al., 2000). However, competences acquired out of school are evaluated with standards of the formal system. To make the discussion even more paradoxical, because the workplace or the NGOs (see recreational activity study book, Finland, 3.7) are such wonderful learning arenas, the learning outcomes have to be valued in order to become eligible for the formal schooling.

There is another threat to this policy. Research shows that job-related continuous learning in non-formal settings is an important aspect for employed people aged 19 to 64 in Germany. In the year 2000, two thirds of the employed said they practised this type of self-education. However, people who failed to complete dual education (blue-collar workers, immigrants and working women) were under-represented (Kuwan

and Thebis, 2001). Findings confirming that some people do not reach the full operating ability during their lifetime provide another example (Oerter, 1980). As a consequence, valuing learning outcomes acquired in non-formal

settings might support the Mathew effect – i.e. giving more to those who already have – and jeopardise fairness if no support grounded in learning and domain specific development theory is added.

List of abbreviations

AERA	American educational research association
IT	Information technology
NVQ	National vocational qualification
PISA	Programme for international student assessment
SVQ	Scottish vocational qualification
TIMSS	Third international mathematics and science study
VET	Vocational education and training

Annex 1: Sample overall training plan (*Bundesanzeiger*, 1998, pp. 8-9)

Lfd. Nr.	Teil des Ausbildungsberufsbildes	Zu vermittelnde Fertigkeiten und Kenntnisse
1	2	3
4.	Geld- und Vermögensanlage (§ 3 Nr. 4)	
4.1	Anlage auf Konten (§ 3 Nr. 4.1)	a) Kunden über Anlagemöglichkeiten auf Konten einschließlich der Sonderformen des ausbildenden Unternehmens beraten b) Konten eröffnen, führen und abschließen c) Kunden über rechtliche Bestimmungen und vertragliche Vereinbarungen informieren d) Kunden über Verfügungsberechtigungen und Vollmachten beraten e) Kunden über Zinsgutschriften und über deren steuerliche Auswirkungen informieren
4.2	Anlage in Wertpapieren (§ 3 Nr. 4.2)	Kunden über Anlagemöglichkeiten, insbesondere in Aktien, Schuldverschreibungen und Investmentzertifikaten, informieren Kunden über rechtliche Bestimmungen und vertragliche Vereinbarungen informieren Chancen und Risiken der Anlage in Wertpapieren einschätzen Kunden über Kursnotierungen und Preisfeststellungen Auskunft geben bei der Abwicklung einer Wertpapierorder mitwirken Kundenanfragen zu Wertpapierabrechnungen beantworten Kunden über Verwahrung und Verwaltung von Wertpapieren beraten Kunden über Ertragsgutschriften und deren steuerliche Auswirkungen informieren Finanzderivate und deren Risiken in Grundzügen beschreiben
4.3	Anlage in anderen Finanzprodukten (§ 3 Nr. 4.3)	a) Vertrieb von Verbundprodukten zur Kapitalanlage und zur Risikovorsorge im Rahmen der Organisation des ausbildenden Unternehmens erklären b) beim Abschluss von Bausparverträgen mitwirken c) Kunden über Möglichkeiten der Kapitalanlage und der Risikovorsorge durch Abschluss von Lebensversicherungen informieren

Annex 2: Sample of skeleton school syllabus (*Bundesanzeiger*, 1998, p. 18)

4. Lernfeld **Geld- und Vermögensanlagen anbieten**

1. Ausbildungsjahr
Zeitrichtwert 100 Stunden

Zielformulierung:

Die Schülerinnen und Schüler ermitteln Bedarfssignale und Anlagemotive der Kunden. Sie präsentieren Finanzinstrumente kundenorientiert. Sie erläutern Preiseinflussfaktoren, Kursbildung und Kursveröffentlichungen. Sie werten Produkt- und Marketinginformationen aus. Sie nutzen produktbezogene Berechnungen. Sie erläutern aus der Anlageentscheidung resultierende Serviceleistungen. Sie beschreiben Risiken, die aus Anlageentscheidungen entstehen, und beachten die Vorschriften des Anlegerschutzes.

Inhalte:

Anlagen auf Konten am Beispiel der Spareinlage: Vertragsgestaltung aus Kunden- und Bankensicht, Bedeutung der Sparurkunde, Regelverfügungen und vorzeitige Verfügungen, Verzinsung, Besteuerung der Zinserträge

Termineinlagen, Sparbriefe

Besonderheiten des Bausparens und der Kapitallebensversicherung gegenüber anderen Anlageformen

Schuldverschreibung, Aktie und Investmentzertifikat als Grundformen der Wertpapiere: Rechtsnatur, Rechte der Inhaber, Ausstattung, Risiken, Emissionsgründe

Kursbildung und Kursnotierung am Beispiel von Aktien; Kurszusätze, Kurshinweise

Grundlagen und Grundsätze der Anlageberatung

Verwahrung und Verwaltung: Girosammelverwahrung, Wertpapierrechnung; Depotstimmrecht

Maßnahmen zum Schutz der Anleger

5. Lernfeld **Besondere Finanzinstrumente anbieten und über Steuern informieren**

2. Ausbildungsjahr
Zeitrichtwert 60 Stunden

Zielformulierung:

Die Schülerinnen und Schüler präsentieren Finanzinstrumente für besondere Anlagewünsche. Sie werten Produkt- und Marktinformationen aus und nutzen produktbezogene Berechnungen. Sie stellen Grundbegriffe des Einkommensteuerrechts und die rechtlichen Rahmenbedingungen zu Geld- und Vermögensanlage dar. Sie geben einen Überblick über die Finanzmärkte und erklären deren einzel- und gesamtwirtschaftliche Bedeutung.

Inhalte:

Wertpapiersonderformen am Beispiel von Genussschein und Optionsanleihe; Rechte der Inhaber, Ausstattung, Risiken, Emissionsgründe

Finanzderivate am Beispiel einer Aktien-Option und eines Futures: Rechte des Inhabers, Risiken, Einsatzmöglichkeiten

Grundbegriffe des Einkommenssteuerrechts

Steuerliche Gesichtspunkte bei der Anlage in Wertpapieren: Besteuerung von Erträgen und Kursgewinnen am Beispiel von Aktien und Schuldverschreibungen

Finanzmärkte: Arten, Funktionen, Bedeutung

Rahmenbedingungen des Kreditwesengesetzes und des Wertpapierhandelsgesetzes zur Geld- und Vermögensanlage

Annex 3: Unit 3 – sell financial products and services (NVQ Level 3, n.d.)

This unit relates to non-mortgage and non-regulated products and services, such as non-regulated insurance products, investments and foreign currency. You will need to show that you can identify your customer's needs and promote suitable products and services by presenting them accurately and seeking your customer's commitment to buy.

Elements

- 3.1. Establish customer needs
- 3.2 Promote the features and benefits of the products and services
- 3.3 Gain customer commitment and evaluate sales techniques

Knowledge and understanding

For each of the elements you must show that you know and understand the following:

- your organisation's requirements relating to relevant codes, laws and regulations for selling products and services
- which products and benefits of products and services you are authorised to promote
- key features and benefits of products and services you are responsible for promoting
- your organisation's sales process relevant to your area of responsibility
- the limits of your authorisation and responsibility when providing information and offering advice on your organisation's products and services
- to whom you should refer customers for information and advice outside of your authorisation and responsibility
- questioning techniques, such as when to use open or closed questions while selling
- the types and forms of buying signals your customers might use
- how to monitor and evaluate your sales technique.

Element 3.1: establish customer needs

Performance criteria

You will need to:

- 3.1.1 Give customers prompt attention and treat them politely
- 3.1.2 Give your customer, when necessary, a clear and accurate account of your role, and the levels of information and advice that you can provide
- 3.1.3 Establish the needs of your customer through discussion, using appropriate questioning
- 3.1.4 Check your customer's responses to ensure you understand them
- 3.1.5 Conduct your discussion in a manner appropriate to your customer's needs
- 3.1.6 Record all relevant information accurately
- 3.1.7 Pass promptly to the relevant person any sales opportunity outside your area of knowledge or responsibility

Evidence requirements

You must be able to demonstrate your competence at establishing customer needs. You must provide evidence from real work activities, undertaken by yourself, that:

- you have established customer needs through:
 - your discussion with the customer
 - your records of these meetings

- you have established needs of customers that are:
 - immediate
 - future.

Examples of evidence

Performance

- Interview/discussion notes and records
- Referral documentation, where sales opportunities were passed on to a relevant colleague

Knowledge and understanding

Make sure you can answer the following questions:

- What are the stages that you follow in selling your organisation's products and services? How might these be adapted to meet the particular needs of different customers?
- What different types of customer needs are there? Why is it important to prioritise these needs when speaking to individual customers?
- In what ways can you check that customers understand the information that you are giving?
- What are the limits of your authority and responsibility when advising customers about your organisation's products and services? What should you do if these limits are reached?

Element 3.2: promote the features and benefits of the products and services

Performance criteria

You will need to:

- 3.2.1 Explain clearly and accurately the features and benefits of products and services that most closely match your customer's needs
- 3.2.2 Identify, issue and explain fully to your customer the promotional material relating to products and services that meet their needs
- 3.2.3 Follow your organisation's procedures to ensure that the options relating to the products and services you offer to your customer conform to relevant codes, and legal and regulatory requirements
- 3.2.4 Address fully and accurately your customer's questions and concerns in a manner that promotes the sale
- 3.2.5 Identify buying signals and correctly act on these
- 3.2.6 Confirm with your customer their understanding of the proposed products and services
- 3.2.7 Identify opportunities for cross-selling products and services and promote these clearly to your customer in a manner that maintains goodwill
- 3.2.8 Pass promptly to the relevant person sales opportunities outside your area of knowledge or responsibility
- 3.2.9 Record all relevant information accurately

Evidence requirements

You must be able to demonstrate your competence at promoting the features and benefits of products and services through:

- your sales performance with customers
- your corresponding customer records.

Examples of evidence

Performance

- Interview/discussion notes and records
- Referral documentation, where sales opportunities were passed to a relevant colleague
- Your customer records

Knowledge and understanding

Make sure you can answer the following questions:

- What is the difference between 'features' and 'benefits'? How does this apply to the different products and services you deal with?
- What are your organisation's procedures and requirements for ensuring that you meet relevant codes, laws and regulations when selling products and services?
- What sort of buying signals might customers use to show their interest, or disinterest, in the products and services being offered? How could you deal with the different types of buying signals?
- What opportunities might you encounter for cross-selling? How would you deal with these?
- What types of sales opportunities are outside your area of responsibility? To whom do you refer them?

Element 3.3: gain customer commitment and evaluate sales technique*Performance criteria*

You will need to:

- 3.3.1 Agree the preferred products and services with your customer
- 3.3.2 Establish and agree the way forward with your customer, and record this accurately
- 3.3.3 Promptly and accurately complete documentation and check that it is signed by your customer in accordance with your organisation's procedures under the relevant codes, and legal and regulatory requirements
- 3.3.4 Deal with documentation in accordance with your organisation's procedures
- 3.3.5 Inform relevant parties of the outcome according to your organisation's procedures
- 3.3.6 Conduct all business in a manner that maintains professional integrity and goodwill
- 3.3.7 Evaluate objectively the effectiveness of the sales technique and use this to influence future sales activity

Evidence requirements

You must be able to demonstrate your competence at achieving a purchase commitment and evaluating your sales technique through:

- your sales performance with customers
- your relevant records

You must provide evidence from real work activities, undertaken by yourself, that:

- you agreed the way forward with customers, to include both:
 - an agreed sale or successful referral with at least one customer
 - sales contact with at least one other
- you can objectively evaluate your own sales technique, through:
 - your records
 - discussions with colleagues and your assessor.

*Examples of evidence**Performance*

- Interview/discussion notes and records
- Product/service application documents completed by yourself and signed by your customer
- Customer records
- Notes/memos notifying the appropriate person(s) of the outcome of your sales discussions
- Notes of your evaluation of your sales technique with proposals for the future.

Knowledge and understanding

Make sure you can answer the following questions:

- What procedures should you follow for completing a sale, including gaining customer commitment, and subsequently completing and forwarding necessary documentation?
- What factors in your sales technique might influence whether a customer agrees to buy from your organisation?
- How could you check the effectiveness of differences in your sales technique?

References

- AAC-project. *Accreditation of achieved competencies*. Leonardo da Vinci project. Wageningen: Stoas Research, 2001.
- Achtenhagen, F. et al. *Social sciences. COST technical committee*. Brussels: European Commission, DG Science, Research and Development, 1995, Vol. 3.
- Achtenhagen, F.; Thang, P. O. *Mehrdimensionale Lehr-Lern-Arrangements*. Wiesbaden: Gabler, 2002, p. 444-459.
- AERA - American educational research association. *AERA position statement concerning high stakes testing in PreK-12 education*. Washington, DC: AERA, 2000. Available from Internet: <http://www.aera.net/about/policy/stakes.htm> [cited 13.2.2004].
- AERA; APA; NCME. *Standards for educational and psychological testing*. Washington, DC: AERA – American Educational Research Association, 1999.
- Alheit, P. *Biographien in der späten Moderne: Verlust oder Wiedergewinn des Sozialen?* Bremen: IBL, 1998. Available from Internet: www.ibl.uni-bremen.de/publik/vortraege/98-01-pa.htm [cited 13.2.2004].
- Anderson, L. W.; Krathwohl, D. R. *A taxonomy for learning, teaching and assessing*. New York: Longman, 2001.
- Anderson, N. Decision making in the graduate selection interview. *Journal of Occupational Psychology*, 1990, No 63, p. 63-76.
- Baker E. L. et al. Policy and validity prospects for performance-based assessment. *American Psychologist*, 1993, No 48, p. 1210-1218.
- Baumert, J. et al. *TIMSS – Mathematisch-naturwissenschaftlicher Unterricht im internationalen Vergleich*. Opladen: Leske und Budrich, 1997.
- Baumert, J. et al. *TIMSS/III – Dritte Internationale Mathematik- und Naturwissenschaftsstudie. Mathematische und naturwissenschaftliche Bildung am Ende der Schullaufbahn*. Opladen: Leske und Budrich, 2000, Vol. 1.
- Beck, K. *Die empirischen Grundlagen der Unterrichtsforschung. Eine kritische Analyse der deskriptiven Leistungsfähigkeit von Beobachtungsmethoden*. Göttingen: Hogrefe, 1987.
- Beck, K.; Krumm, V. *Wirtschaftskundlicher Bildungs-Test*. Göttingen: Hogrefe, 1998.
- Berufsbildungsbericht 2003. In: *BMBF Publik.* Bonn: Bundesministerium für Bildung und Forschung (ed.), 2003.
- Björnavold, J. *Making learning visible: identification, assessment and recognition of non-formal learning in Europe*. Luxembourg: Office for Official Publication of the European Communities, 2000 (Cedefop Reference series).
- Björnavold, J.; Tissot, P. Glossary. In: Björnavold, J. *Making learning visible: identification, assessment and recognition of non-formal learning in Europe*. Luxembourg: Office for Official Publication of the European Communities, 2000, p. 199-221 (Cedefop Reference series).
- BMBF. *IT-Weiterbildung mit System. Neue Perspektiven für Fachkräfte und Unternehmen*. Bonn: BMBF, 2002.
- Bloom, B. S. (ed) *Taxonomy of educational objectives: handbook I: cognitive domain*. New York: Adavid McKay, 1956.
- Boyatzis, R. E. *The competent manager: a model for effective performance*. New York: Wiley, 1982.
- Bransford, J. D. et al. *How people learn*. Washington, DC: National Academic Press, 2000.
- Brown, S. *Fit for purpose assessment*. Paper presented at the EARLI conference, Gothenburg, August 1999.
- Bulmahn, E.; Wolff, K.; Klieme, E. *Zur Entwicklung nationaler Bildungsstandards: eine Expertise*. Frankfurt a.M.: IPF/BMBF, 2003. Available from Internet: http://www.dipf.de/aktuelles/expertise_bildungsstandards.pdf [cited 16.2.2004].
- Bundesanzeiger. Bekanntmachung der Verordnung über die Berufsausbildung zum Bankkaufmann/zur Bankkauffrau nebst Rahmenlehrplan vom 23. Januar 1998. *Bundesanzeiger*, 1998, Vol. 50, No 82a.
- Campbell, D. T.; Stanley, J. C. Experimental and quasi-experimental designs for research on teaching. In: *Handbook of research on teaching*. Skokie, Ill: Rand McNally, 1963, p. 171-246.
- Cedefop. *Vocational education and training – the European research field: background report*

- (two volumes). Luxembourg: Office for Official Publication of the European Communities, 1998 (Cedefop Reference series).
- Chomsky, N. Rules and representations. *The behavioural and brain sciences*, 1980, Vol. 3, p. 1-61.
- Cseh, M. et al. Informal and incidental learning in the workplace. In: Straka, G. A. (ed.) *Conceptions of self-directed learning*. Münster: Waxmann, 2000, p. 59-74.
- CTGV. *The Jasper project*. The cognition and technology group at Vanderbilt. Mahwah, NJ: Lawrence Erlbaum, 1997.
- Deutsches PISA-Konsortium (eds). *PISA 2000 – Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich*. Opladen: Leske und Budrich, 2001.
- Deutsches PISA-Konsortium. *Die Bundesländer der Bundesrepublik Deutschland im Vergleich*. Opladen: Leske und Budrich, 2002.
- DIHT. *Abschlussprüfung Bankkaufmann/Bankkauffrau: Ausbildungsordnung vom 30. Dezember 1997: Prüfungsfach 'Kundenberatung'*. Bonn: Deutscher Industrie- und Handelstag, 1999.
- DIN. *DIN 33430: Anforderungen an Verfahren und deren Einsatz bei berufsbezogenen Eignungsbeurteilungen [job related proficiency assessment]*. Berlin: Deutsches Institut für Normung, 2002.
- Dörner, D. *Problemlösen als Informationsverarbeitung*. Stuttgart: Kohlhammer, 1976.
- Embretson, S. E.; Reise, S. P. *Item response theory for psychologists*. Mahwah: Lawrence Erlbaum, 2000.
- Eraut, M. *Developing professional knowledge and competence*. London: Taylor and Francis, 1994.
- Eraut, M. Non-formal learning and tacit knowledge in professional work. *British Journal of Educational Psychology*, 2000, No 70, p. 113-136.
- Eraut, M. The role and use of vocational qualifications. *National Institute Economic Review*, 2001, No 178, p. 88-98.
- Eraut, M. Editorial. *Learning in health and social care*, 2003a, p. 1-3.
- Eraut, M. National vocational qualifications in England: description and analysis of an alternative qualification system. In: Straka, G. A. (ed.) *Zertifizierung non-formell und informell erworbener beruflicher Kompetenzen*. Münster: Waxmann, 2003b, p. 117-123.
- Eraut, M. et al. *Evaluation of higher level S/NVQs*. Sussex: University, 2001 (Research report).
- Ertl, H. *Modularisation of vocational education in Europe: NVQs and GNVQs as a model for the reform of initial training provisions in Germany?* Wallingford: Symposium Books, 2000.
- Ertl, H. *The role of EU programmes and approaches to modularisation in vocational education: fragmentation or integration?* München: Herbert Utz Verlag, 2002.
- European Commission. *A memorandum on lifelong learning*. Brussels: European Commission – DG for Education and Culture, 2000 (SEC (2000) 1832).
- European Commission. *Making a European area of lifelong learning a reality: communication from the Commission*. Luxembourg: Office for Official Publication of the European Communities, 2001 (Documents COM, (2001) 678).
- Gagné, R. M. *The conditions of learning*. New York: Holt, Rinehart and Winston, 1977.
- Gilbert P. *L'évaluation des compétences à l'épreuve des faits*. Luxembourg: Entreprise et Personnel, 1998.
- Grant, G. et al. *On competence: a critical analysis of competence-based reforms in higher education*. San Francisco: Jossey-Bass, 1979.
- Grunwald, S. Zertifizierung arbeitsprozessorientierter Weiterbildung. In: Rohs, M. (ed.) *Arbeitsprozessintegriertes Lernen: neue Ansätze für die berufliche Bildung*. Münster: Waxmann, 2002, p. 165-179.
- Grunwald, S.; Rohs, M.. Zertifizierung informell erworbener Kompetenzen im Rahmen des IT-Weiterbildungssystems. In: Straka, G. A. (ed.) *Zertifizierung non-formell und informell erworbener beruflicher Kompetenzen*. Münster: Waxmann, 2003, p. 207-221.
- Gutschow, K. Erfassen, Beurteilen und Zertifizieren non-formell und informell erworbener beruflicher Kompetenzen in Frankreich. In: Straka, G. A. (ed.) *Zertifizierung non-formell und informell erworbener beruflicher Kompetenzen*. Münster: Waxmann, 2003, p. 127-139.
- Kerlinger, F. N. *Foundations of behavioural research*. London: William Clowes, 1973.
- Klauer, K. J. *Revision des Erziehungsbegriffs*. Düsseldorf: Schwann, 1973.
- Klieme, E. et al. Mathematische und naturwissenschaftliche Grundbildung: Konzeptuelle Grundlagen und die Erfassung und Skalierung

- von Kompetenzen. In: Baumert, J. et al. *TIMSS/III*, Vol. 1. Opladen: Leske und Budrich, 2000, p. 85-133.
- KMK. *Handreichungen für die Erarbeitung von Rahmenplänen der KMK für den berufsbezogenen Unterricht in der Berufsschule und ihre Abstimmung mit Ausbildungsordnungen des Bundes für anerkannte Ausbildungsberufe*. Sekretariat der ständigen Konferenz der Kultusminister der Länder in der BRD, 1996/2000.
- Krumm, V.; Weiß, S. Ungerechte Lehrer. *Psychozial*, 2000, Vol. 23, No 79, p. 57-73.
- Kuwan, H.; Thebis, F. *Berichtssystem Weiterbildung VIII: Erste Ergebnisse der Repräsentativbefragung zur Weiterbildungssituation in Deutschland*. Bonn: Bundesministerium für Bildung und Forschung, 2001.
- Lienert, G. A.; Raatz, U. *Testaufbau und Testanalyse*. München: Beltz Psychologie Verlags Union, 1998.
- Linn, R. L. Assessments and accountability. *Educational Researcher*, 2000, Vol. 29, No 2, p. 4-16.
- Linn, R. L.; Baker, E. L. Comparing results from disparate assessments. *CRESST Line*, Winter 1993, p. 1-3.
- Linn, R. L.; Gronlund, N. E. *Measurement and assessment teaching*. New Jersey: Prentice Hall, 2000.
- Ministry of Education. *DACUM booklets*. British Columbia: Ministry of Education, 1983.
- Nijhof, W. J. Assessing competencies at work – reflections on a case in the Netherlands. In: Straka, G. A. (ed.) *Zertifizierung non-formell und informell erworbener beruflicher Kompetenzen*. Münster: Waxmann, 2003, p. 141-151.
- Norris, N. The trouble with competence. *Cambridge Journal of Education*, 1991, Vol. 21, No 3, p. 331-341.
- NRC – National Research Council. *Knowing what students know: the science and design of educational assessments*. Committee on the Foundations of Assessment. Washington, DC: National Academies Press, 2001.
- NVQ. *Providing financial services (banks and building societies)*. NVQ Level 2. Standards and assessment guidance. The chartered institute of bankers. n.d.
- NVQ. *Providing financial services (banks and building societies)*. NVQ Level 3. Standards and assessment guidance. The chartered institute of bankers. n.d.
- CIB. *Providing financial services (banks and building societies)*. NVQ Level 2. Standards and assessment guidance. The chartered institute of bankers, n.d.
- CIB. *Providing financial services (banks and building societies)*. NVQ Level 3. Standards and assessment guidance. The chartered institute of bankers, n.d.
- OECD. *Knowledge and skills for life – first results from PISA 2000*. Paris: OECD, 2001.
- Oerter, R. *Moderne Entwicklungspsychologie*. Donauwörth: Auer, 1980.
- O'Grady, M. Assessment of prior achievement/assessment of prior learning: issues of assessment and accreditation. In: *The Vocational Aspect of Education*, 1991, No 115.
- Pettersen, I. L. (ed.) *Competence card from the workplace. A cooperation between trade and industry and the educational system in Nordland*, n.d.
- Pettersen, I. L. (ed.) *CV. Curriculum Vitae. A cooperation between trade and industry and the educational system in Nordland*, n.d.
- Polanyi, M. *The tacit dimension*. Garden City, NY: Doubleday, 1966.
- Popper, K. R. *Logik der Forschung*. Tübingen: Mohr, 1989.
- Porter, M. E. *Competitive advantage*. New York: Macmillan, 1985.
- Pütz, H. 'Berufsbildungs-PISA' wäre nützlich. *Berufsbildung in Wissenschaft und Praxis*, 2002, Vol. 31, No 3, p. 34.
- QCA. *Assessing NVQs. March 1998, QCA Guarding Standards*. London: QCA – Qualifications and Curriculum Authority, 1998.
- QCA. *Developing an assessment strategy for NVQs and SVQs*. London: QCA – Qualifications and Curriculum Authority, 1999 (Ref: QCA/99/396).
- Resnick, L. B. Learning in school and out. *Educational Researcher*, 1987, Vol. 12, p. 13-20.
- Savisaari, L. *Youth academy & recreational activity study book system*. Helsinki: Nuorten Akatemia, 2002. Available from Internet: http://www.nuortenakatemia.fi/index.php?lk_id=64 [cited: 17.2.2004].
- Savisaari, L. The recreational activity study book system as a tool for certification of youth non-formal and informal learning in Finland – a practitioner's approach. In: Straka, G. A. (ed.) *Zertifizierung non-formell und informell erwor-*

- bener beruflicher Kompetenzen. Münster: Waxmann, 2003, p. 223-229.
- Schuler, H. *Psychologische Personalauswahl*. Göttingen: Verlag für Angewandte Psychologie, 2000.
- Semill, D. *Problemlösefähigkeit, Handlungskompetenz und emotionale Befindlichkeit*. Göttingen: Verlag für Psychologie, 1992.
- Shepard, L. A. The role of assessment in a learning culture. *Educational Researcher*, 2000, Vol. 29, No 7, p. 4-14.
- Shepard, L. A. The role of classroom assessment in teaching and learning. In: Richardson, V. (ed.) *Handbook of research on teaching* (fourth edition). Washington, DC: AERA – American Educational Research Association, 2001, p. 1066-1101.
- Sternberg, R. J.; Kolligian, J. Jr. *Competence considered*. New Haven: Yale University Press, 1990.
- Straka, G. A. *Forschungsstrategien zur Evaluation von Schulversuchen*. Weinheim: Beltz, 1974.
- Straka, G. A. (ed.) *European views of self-directed learning*. Münster: Waxmann, 1997.
- Straka, G. A. (ed.) *Conceptions of self-directed learning*. Münster: Waxmann, 2000.
- Straka, G. A. Action-taking orientation of vocational education and training in Germany – a dynamic concept or rhetoric? In: Nieuwenhuis, L. F. M.; Nijhof, W. J. (eds) *The dynamics of VET and HRD systems*. Enschede: Twente University Press, 2001a, p. 89-99.
- Straka, G. A. Lern-lehr-theoretische Grundlagen der beruflichen Bildung. In: Bonz, B. (ed.) *Didaktik der beruflichen Bildung*. Baltmannsweiler: Schneider Verlag, 2001b, p. 6-30.
- Straka, G. A. Valuing learning outcomes acquired in non-formal settings. In: Nijhof, W. J.; Heikkinen, A.; Nieuwenhuis, L. F. M. (eds) *Shaping flexibility in vocational education and training*. Dordrecht: Kluwer, 2002a, p. 149-165.
- Straka, G. A. Empirical comparisons between the English national vocational qualifications and the educational and training goals of the German dual system – an analysis for the banking sector. In: Achtenhagen, F.; Thång, P.-O. (eds) *Transferability, flexibility and mobility as targets of vocational education and training*. Proceedings of the final conference of the COST Action A11, Gothenburg, 13 and 16 June, 2002b, p. 227-240.
- Straka, G. A. Rituale zur Zertifizierung formell, non- und informell erworbener Kompetenzen. *Wirtschaft und Erziehung*, 2003a, Vol. 55 (7-8), p. 253-259.
- Straka, G. A. (ed.) *Zertifizierung non-formell und informell erworbener beruflicher Kompetenzen*. Münster: Waxmann, 2003b.
- Straka, G. A. *Verfahren und Instrumente der Kompetenzdiagnostik – der Engpass für ein Berufsbildung-PISA?* (In preparation).
- Straka, G. A.; Lenz, K. Was trägt zur Entwicklung von Fachkompetenz in der Berufsausbildung bei? *Schweizerische Zeitschrift für kaufmännisches Bildungswesen*, 2003, Vol. 97, p. 54-67.
- Straka, G. A.; Macke, G. *Lern-Lehr-Theoretische Didaktik*. Münster: Waxmann, 2002.
- Straka, G. A.; Macke, G. Handlungskompetenz und Handlungsorientierung als Bildungsauftrag der Berufsschule – Ziel und Weg des Lernens in der Berufsschule? *Berufsbildung in Wissenschaft und Praxis*, 2003, Vol. 32, No 4, p. 43-47.
- Straka, G. A.; Schaefer, C. *Validating a more-dimensional conception of self-directed learning*. Paper presented at the international research conference of the AHRD, in Honolulu, Hawaii, from 27 February to 3 March 2002.
- Thommen, J.-P. *Allgemeine Betriebswirtschaftslehre. Umfassende Einführung aus managementorientierter Sicht*. Wiesbaden: Gabler, 1991.
- Thömmes, J. Bilan de compétences. In: Erpenbeck, J.; von Rosenstiel, L. (eds) *Handbuch Kompetenzmessung*. Stuttgart: Schäffer-Poeschel, 2003, p. 545-555.
- Ulrich, H. *Die Unternehmung als produktives soziales System*. Bern: Haupt, 1970.
- VOX. *The Realkompetanse Project*. Validation of non-formal and informal learning in Norway. Oslo: VOX, 2002.
- Watermann, R.; Baumert, J. Mathematische und naturwissenschaftliche Grundbildung beim Übergang von der Schule in den Beruf. In: Baumert, J.; Bos, W.; Lehmann, R. (eds) *TIMSS/III – Mathematische und naturwissenschaftliche Bildung am Ende der Schullaufbahn*. Opladen: Leske und Budrich, 2000, Vol. 1, p. 199-259.
- Weinert, F. E. Concept of competence: a conceptual clarification. In: Rychen, D. S.; Salganik, L. H. (eds) *Defining and selecting key competencies*. Göttingen: Hogrefe, 2001, p. 45-66.

- Weinert, F. E. *Concepts of competence*. München: Max Planck Institut, 1999. Available from Internet: http://www.statistik.admin.ch/stat_ch/ber15/deseco/weinert_report.pdf [cited: 18.2.2004]
- Weisschuh, B. Nachhaltige Qualitätssicherung bei der Beurteilung und Förderung von Schlüsselqualifikationen im Produktionsbetrieb. In: Straka, G. A. (ed.) *Zertifizierung non-formell und informell erworbener beruflicher Kompetenzen*. Münster: Waxmann, 2003, p. 231-242.
- Wöhe, Günter. *Einführung in die Allgemeine Betriebswirtschaftslehre*. München: Vahlen, 2002.
- Wolf, A. *Competence-based assessment*. Buckingham: Open University Press, 1995.
- Wolf, A. Competence-based assessment. Does it shift the demarcation line? In: Nijhof, W. J.; Streumer, J. N. (eds) *Key qualifications in work and education*. Dordrecht: Kluwer Academic Press, 1998, p. 207-220.



An overarching conceptual framework for assessing key competences in an international context

Lessons from an interdisciplinary and policy-oriented approach

Dominique Simone Rychen

Abstract

The OECD project DeSeCo (Definition and selection of competences: theoretical and conceptual foundations) was undertaken in response to this increasing interest in education outcomes and their effects. DeSeCo developed a common, overarching conceptual frame of reference for identifying and assessing key competences. Key competences are individually based competences considered necessary or desirable for effective participation in democratic societies and for coping with global demands, particularly those related to the so-called knowledge economy or information society.

DeSeCo was initiated in the OECD context at the end of 1997 and carried out under the leadership of the Swiss Federal Statistical Office. It is embedded in OECD's long-term programme on education indicators (INES) which aims to provide measures on the functioning, development and impact of education. The work of DeSeCo was designed to complement past and current international empirical studies, in particular the International Adult Literacy Survey (IALS), the Programme for International Student Assessment (PISA) and the Adult Literacy and Life Skills (ALL) survey. The analysis and reflection in DeSeCo is not restricted to what can be learned and taught in schools or to what is currently or readily measurable in large-scale assessments. DeSeCo has not addressed its task by an inductive method, starting from factual situations, but rather by starting at a more general level, laying out conceptual and theoretical considerations.

DeSeCo's approach is international, interdisciplinary and policy-oriented. Inputs to DeSeCo have included discipline-oriented academic contributions, commentaries from leading experts working in various fields, as well as country-specific reports. Two international symposia and various expert meetings provided opportunities for gaining interdisciplinary insight and a better understanding of the needs and concerns of policy-makers and experts from education, business, labour, health and other relevant sectors. These varied inputs have led to the construction of an overarching conceptual frame of reference for assessing and developing key competencies.



Table of contents

1. Introduction	315
2. Background information	316
2.1. Policy context	316
2.2. Research context and focus of the study	316
3. Necessary steps in the definition and selection of key competences	318
3.1. Analysis of international comparative studies on learning outcomes	318
3.2. Towards a clarification of the concept of competence	318
3.3. Expert opinions from multiple disciplines and sectors	319
3.4. Country contribution process	319
3.5. Debates and exchange: from multiple perspectives towards interdisciplinary insight	319
4. Elements of DeSeCo's overarching frame of reference for key competences	321
4.1. Criteria for competences and key competences	321
4.1.1. The concept of competence	321
4.1.2. The concept of key competence	321
4.2. Three categories of key competences	322
4.2.1. Acting autonomously	322
4.2.2. Using tools interactively	323
4.2.3. Interacting in socially heterogeneous groups	323
4.3. Interacting of key competences	323
5. Implications for a coherent assessment strategy	324
5.1. Dealing with the complex nature of key competences	324
5.2. Towards a coherent assessment strategy	325
6. Conclusions	326
List of abbreviations	327
Bibliography	328

1. Introduction

The DeSeCo project (Definition and selection of competences: theoretical and conceptual foundations) is a response to several trends that arose independently in different sectors that are now beginning to converge, with major policy implications. One trend is a concern for people's abilities to cope with the many rapid changes that are taking place in the world, including the emergence of the so-called knowledge economy and information society. Although there is little consensus on specific competences and skills that individuals will need, there is considerably more consensus on more general requirements, such as creativity, initiative taking and adaptability. Another trend is the shift from input- to outcome-based public policies, particularly in education and training and the resulting demand for relevant statistical data and indicators. Simultaneously, large-scale educational assessments are expanding from measuring academic achievement alone to considering individuals' ability to apply knowledge and skills in non-school contexts. This expansion brings with it a number of new terms and concepts, but little consistency in their usage.

In light of these trends, an attempt to relate the demands of the modern world to abilities, skills and other qualities individuals need in order to

provide useful information for policy and planning is both appealing and logical. In the context of the OECD, the natural starting point was its system of education-related indicators. Although these indicators have undergone significant development and have always been policy-relevant, a broader perspective on what should be assessed and an overarching theoretical framework for such assessments were deemed necessary to help guide future indicator development and long-term assessment strategies.

Throughout its brief history, DeSeCo has sought to include a broad range of perspectives not only from the academic world, but from the worlds of public policy, labour, education and business as well. Attention has also been given to the varying contexts found throughout the participating countries. Each input has contributed a unique combination of perspective, experience, expertise and interest that has helped shape the goals, definitions, limitations and potential of DeSeCo. By taking an approach that was both conceptual and pragmatic, it is hoped that DeSeCo will enhance the relevance of international assessments. Based on the outcomes of this study we are now prepared to establish and affirm several important guiding principles and discuss the implications of DeSeCo for further work.

2. Background information

2.1. Policy context

In light of common challenges and transformations taking place in economies and societies throughout the world, education is widely viewed as an important investment and asset for both the individual and the community. While economic growth remains a main objective for governments and societies, there are increasing policy concerns about its impact on the natural and social environments (for instance OECD, 2001a). Lifelong learning and competence or skill development have become key notions in international efforts to enhance economic and social development (for instance the World Bank, 2002; OECD, 2001b). With this heightened importance placed on education and lifelong learning, tomorrow's curriculum and educational goals have become a relevant topic in political discourses all over the world (within OECD see for instance Trier, 2003). There is a growing concern among governments, employers and the general public about the adequacy and quality of education and training. This concern is, for example, reflected in the European Commission communication, *Making a European area of lifelong learning a reality* (2001, p. 3): 'Traditional policies and institutions are increasingly ill-equipped to empower citizens for actively dealing with the consequences of globalisation, demographic change, digital technology and environmental damage. Yet people, their knowledge and competences are the key to Europe's future'.

With a shift from input- to outcome-oriented policies, the OECD and other international institutions have invested considerable effort in developing comparable outcome indicators in the education field to provide information relevant for government policy (Salganik et al., 1999; Salganik, 2001). In general, these indicators reflect traditional notions of academic achievement and skill development, such as reading and mathematics skills, either in curriculum-based or real-world-based contexts. This focus partially results from practical considerations, but also

results from the fact that these skills are recognised as foundation skills, crucial to success in the so-called knowledge and information society of today. Though traditional competences such as reading literacy have been measured throughout OECD countries and beyond, there is a broad consensus that we are far from assessing a truly comprehensive set of competences relevant to human and social development and political and economic governance. OECD constituencies, recognising the need for a conceptual basis for defining and selecting relevant competences, thus welcomed the work programme of the DeSeCo project. DeSeCo's theory-based and policy-oriented work programme found broad support and interest within and outside of the OECD.

2.2. Research context and focus of the study

The DeSeCo project originated in a governmental context, the OECD, at the end of 1997 and was carried out under the leadership of the Swiss Federal Statistical Office. It is embedded in OECD's long-term programme on education indicators (INES), aimed at providing measures on the functioning, development and impact of education from early childhood, through formal education and to learning and training throughout life. The work of DeSeCo was designed to complement past and current international empirical studies, in particular the International Adult Literacy Survey (IALS), the Programme for International Student Assessment (PISA) and the Adult Literacy and Life Skills (ALL) survey. Throughout the work programme, close ties with these and other international surveys have been maintained in recognition that success in this complex field depends not only on theoretical and empirical work, but also on the iterative process entailed by constant dialogue and exchange among the various specialists and stakeholders.

The main goal of DeSeCo was the construction of an overarching conceptual frame of reference relevant both for the development of key competences in a lifelong learning perspective and for the assessment of these competences in an international setting. This study does not address the topic of education outcomes by proceeding with an inductive method, starting from factual situations, but rather by starting at a more general level, laying out conceptual and theoretical considerations. DeSeCo considers the topic of important, necessary or desirable competences from a broad, holistic and interdisciplinary perspective. Thus, the reflection on and analysis of key competences is not limited to the school context, student achievement or workers' skills and the demands of the labour market.

While recognising the value and importance of domain- or job-specific competences and technical skills, DeSeCo focuses on competences that apply across multiple areas of life (for instance, in the family, at work, in the political sphere, in the health sector, etc.) and that contribute to personal, economic and social well-being. The leading theme is: in light of the demands of democratic societies and the global challenges of modern life including those related to the knowledge-based economy or information society, what competences are important or necessary for everyone; and, what are the normative assumptions and conceptual criteria through which these competences are constructed?

Considerations of whether these competences can be assessed or measured in large-scale or qualitative studies – and if so, how – were not at the forefront. The determination of potentially 'new' competences is based, instead, on several issues that have not played prominent roles in the past, including:

- (a) theoretical models and concepts – what is a competence? What makes a competence 'key'?
- (b) cultural context, biographical variability – to what extent are the competences identified relevant across cultures and to individuals occupying different positions in society and the economy?
- (c) political negotiation, consensus formation – which competences do policy-makers value and which do they have interest in assessing?
- (d) visions of society and individuals – to what extent do the identified competences depend on a particular set of assumptions about the way the world should be and the way people should behave?

DeSeCo's approach of asking these types of questions, of stepping back from the immediate development of assessment instruments and of providing a platform for reflection and exchange among scholars, policy-makers and experts from various fields proved a useful model for furthering the conceptual and theoretical foundations of relevant competences.

3. Necessary steps in the definition and selection of key competences

The conclusions and recommendations drawn from this interdisciplinary and policy-driven research are the results of a collaborative effort among scholars from different disciplines; experts from education, business, labour, health and other relevant sectors; as well as various stakeholders at the national and international level.

The work programme consists of four main studies and two international symposia. Throughout the project, attention has been given to ensure inclusion of both theoretical considerations, as represented in many of the expert papers (Rychen and Salganik, 2001) and practical and policy-oriented needs, as described in the reports stemming from the country contribution process (Trier, 2003). Since these activities constitute the key sources and reference materials for the construction of an overarching frame of reference and for the conceptualisation of key competences, the following sections will highlight the rationale and main findings for each activity.

3.1. Analysis of international comparative studies on learning outcomes

To understand the state of the field, uncover any gaps in knowledge and work out the most promising strategy at the international level, DeSeCo began in 1998 with an analysis of three major studies on competences that had been conducted within the OECD context: the Cross-Curricular Competences Project, an explorative study; the International Adult Literacy Survey (IALS) and the Human Capital Indicators Project (Salganik et al., 1999). Several central themes were found to transcend these three studies, including: desired outcomes of education are broader than the acquisition of the subject-related knowledge typically taught in school; competences develop beyond the school context throughout life; and education is not an end in itself, but a means for making students competent and thereby prepared for life.

The authors also found that, despite common goals and a connection to the OECD, there were few explicit links between the three studies, in terms of both an overarching theory-grounded framework and a larger coordinated research effort. Since the studies arose from different purposes and focused on different types of population groups, this was neither surprising nor considered a shortcoming of their designs at that time. In all of these studies, pragmatic considerations dominated and, ultimately, priority was given to empirical testing of concepts by applying existing methodologies, instruments and, when possible, existing data. Due to the emphasis on the feasibility of measurement, theoretical and conceptual considerations such as an understanding of the nature of competences and the interrelations among various competences were not an issue. These findings confirmed the need to further advance the theoretical underpinning of competences, in particular by developing an explicit, overarching conceptual framework to guide future work.

More recent and partly still continuing projects, such as the PISA, the ALL survey and the IEA Civic Education Study (CivED), were further analysed in 2000. Two of the main findings of this review, specifically that there has been considerable investment in ‘conceptual work and [the] development of measures grounded in well-established theory’ (Salganik, 2001, p. 30) and that lessons learned from previous studies were more fully incorporated in the new studies, indicate potential interest in and application for DeSeCo’s work.

3.2. Towards a clarification of the concept of competence

In light of the terminological and conceptual confusion associated with notions such as competence, skills, qualifications, standards, literacy and so on, an expert paper was commissioned to provide a preliminary clarification of the

concepts related to competence and key competence (Weinert, 2001). With its systematic analyses of existing terms and the theoretical and conceptual approaches within the social sciences to the concept of competence, the paper served a necessary and important function. In line with Weinert's recommendations as confirmed by subsequent discussions on conceptual issues, DeSeCo, as discussed below, opted for a functional (demand-oriented) approach to competences.

3.3. Expert opinions from multiple disciplines and sectors

In 1999, renowned scholars from different academic disciplines (anthropology, psychology, economics, sociology and philosophy) were asked to construct a set of relevant key competences from their own theoretical background and disciplinary perspective. The authors were expected to justify their selections theoretically, taking into account any available state-of-the-art research-based evidence.

The philosophers Canto-Sperber and Dupuy (2001) drew on their expertise in moral and social philosophy and the philosophy of mind not only to address the question of key competences but also to consider what constitutes 'the good life'. Haste (2001) emphasised from her sociopsychological viewpoint the need to look at individuals in a cultural, social and linguistic context. For Perrenoud (2001), a sociologist, the central question to be addressed was what competences are needed by everyone to freely exercise his or her autonomy in multiple social fields. He associated a successful life with not being abused, alienated, dominated or exploited. Levy and Murnane (2001), as well as Goody (2001), differed from the other authors in so far as they used an empirical approach to the issue of competence. Levy and Murnane used relevant economic theories and empirical results to address the question of the competences workers need to succeed in the labour market. Goody, the anthropologist, rejected engaging in a decontextualised discussion of key competences on the grounds that theory must always be considered in relation to practice.

Subsequently a commenting process was initiated (Rychen and Salganik, 2001), wherein the

scholarly papers were distributed among the authors, other academics and leading representatives from various social sectors. It was a first step to explore main convergences and divergences among the different disciplinary perspectives and to gain insight on priority areas from practical and policy-oriented perspectives. The results of this commenting process constituted a starting point for further exploring common or complementary concepts and models relevant from an interdisciplinary and international perspective.

3.4. Country contribution process

Recognising that defining key competences is a process that draws not only on a scientific foundation but also on negotiations among various stakeholders in the political and economic arenas, a country consultation process was launched to gain insight on the public debate and the education needs and priorities of various national contexts and sectors both within and across countries. A summary report was prepared based on the national contributions submitted from 12 OECD countries (Trier, 2003). Convergence rather than divergence was found across countries and between the economic and education sectors within countries. Similar competence areas with almost identical content are emphasised, even if they are worded differently. Indeed, there is a broad consensus on the importance of social competences, communication, literacies, lifelong learning, personal competences and competences necessary for participation in political life. Value orientation is another aspect of particular concern in several countries.

3.5. Debates and exchange: from multiple perspectives towards interdisciplinary insight

Each of the four activities described above raised specific challenges that DeSeCo had to overcome and emphasised the need to bring together the knowledge and insights gained through these activities. Two international symposia and various expert meetings provided opportunities for moving

from the specific activities and multiple perspectives towards increased understanding and awareness of the issues at stake and eventually to a consensus on an integrated frame of reference.

The first international symposium in 1999 was a forum bringing together scholars and experts, including the authors of the papers (Rychen and Salganik, 2001). It succeeded in creating an international network of interested academics and experts, increasing the awareness of the issues inherent in dealing with questions of competences and their assessment and clarifying and prioritising avenues for further work in this area.

The second international symposium (Rychen et al., 2003), with approximately 170 participants, including policy-makers, policy analysts, assessment specialists, academics and representatives from national and international institutions, provided a forum for further consolidating concepts and ideas on key competences and working towards consensus around the question of what competences are relevant for today's societies and economies. It succeeded in outlining options and strategies for the eventual development of a coherent international assessment programme (OECD, 2002).

4. Elements of DeSeCo's overarching frame of reference for key competences

Given the initial finding of a lack of a common conceptual framework for understanding competences (Salganik et al., 1999; Weinert, 2001), the research conducted within the scope of DeSeCo resulted in a comprehensive frame of reference to situate current and future key competencies in a larger conceptual context. It includes, as core elements, the concepts of competence and key competence and a threefold categorisation for key competences considered as critical for a successful life and a well-functioning society.

4.1. Criteria for competences and key competences

4.1.1. The concept of competence

A competence is defined as the ability to meet a complex demand. Each competence corresponds to a combination of interrelated cognitive and practical skills, knowledge and personal qualities such as motivation, values and ethics, attitudes and emotions. These components are mobilised together for effective action in a particular context. This definition represents a demand-oriented or functional approach, placing at the forefront the manifold demands individuals encounter in the context of work and everyday life. It is holistic, in the sense that it integrates and relates demands, individual attributes and context as integral elements of competent performance.

To facilitate the further discussion of competence, it is important to emphasise that in the DeSeCo project, the terms 'competence' and 'skill' are not used as synonyms. The term 'competence' (a holistic concept) designates a complex action system encompassing knowledge, cognitive skills, attitudes and other non-cognitive components, while the term 'skill' is used to designate an ability to perform motor and/or cognitive acts.

4.1.2. The concept of key competence

Based on the concept of competence described above, one could imagine any number of compe-

tences. DeSeCo's policy orientation, however, requires further focus on key competences in the sense of 'competences of prime importance for a successful life and effective participation in different fields of life – including economic, political, social and family domains; public and private interpersonal relations; and individual personal development' (SFSO, 2001b). This approach is consistent with the general commitment of OECD countries to expand opportunities for individuals in various spheres of life, to improve overall living conditions in society and to invest in the development of competences for all (OECD, 2001b). Based upon a review of existing work in the area of competence and on the input of experts from a variety of backgrounds, two important criteria for key competences are put forth; namely, that key competences:

- (a) contribute to broad desired outcomes of an overall successful life and a well-functioning society;
- (b) enable all individuals to cope successfully with complex demands and challenges across a wide spectrum of relevant contexts and domains.

The first criterion raises the important questions 'What constitutes a successful life?' and, 'What are desirable goals for societies?'. Individual and societal ideals and goals are as varied as individuals and societies themselves. However, numerous international conventions and agreements, such as the Universal Declaration of Human Rights, the Rio Declaration on Environment and the World Declaration on Education for All, exhibit consensus around broad desirable goals for all individuals and countries that can serve as the basis for a common normative framework. For example, the World Declaration on Education for All (1990) described several goals of education that are relevant for all individuals in all societies: 'To survive, to develop their full capacities, to live and work in dignity, to participate fully in development, to improve the quality of their lives, to make informed decisions and to continue learning' (Article 1). Thus, basic principles of human rights, democratic values and

the broad goals of sustainable development provide a normative anchoring point for the discourse on key competences and their selection and development. While more specific normative frameworks could be developed for specific individuals and societies, this framework is considered applicable to all and is specific enough to ensure that key competences relate only to demands that are meaningful and desirable.

The second criterion narrows selection to only those competences that are transversal in the sense that they are necessary for understanding and acting effectively in and across different relevant domains or areas of life such as the workplace, personal life, the health field, the political domain, etc. Thus, competences that are domain-specific, that is, applicable in only one sector of the economy or society, are not deemed key competences. DeSeCo's focus on key competences does not question the relevance of such domain specific competences. These are necessary and often constitute important resources for coping with the demands of particular contexts and situations.

In addition, conceptualising key competences requires consideration of demands and challenges faced by individuals and societies. When speaking of the complexity and challenges of today's world, there is broad consensus that basic skills, although important, are not sufficient. Many scholars and experts agree that effective and responsible participation in today's world calls for the development of a higher mental level of complexity (Kegan, 2001) that implies critical thinking and a reflective, holistic approach to life on the part of the individual. This notion of reflectivity (though a variety of terms are used) is a common thread throughout the various contributions from scholars and experts prepared for DeSeCo (in particular, Kegan, 2001; Canto-Sperber and Dupuy, 2001; Perrenoud, 2001). This higher level of mental complexity (or, using other terms, a critical stance and a reflective/integrated/holistic approach to life) represents a central transversal feature of the identified key competences.

To summarise, key competences as adopted in the OECD are competences (each of which consists of a cluster of cognitive skills and non-cognitive dimensions such as attitudes, motivation, values and emotion, the structure of

which are defined by the demand in question) that help individuals cope with complex demands and thus require critical thinking skills and a reflective/holistic approach to life. Key competences apply to various relevant areas of life and, as such, can be considered transversal competences. They contribute to highly valued individual and societal outcomes (i.e. an overall successful life for individuals and a well-functioning society) consistent with the universal goals and values identified by international conventions.

4.2. Three categories of key competences

Based on the inputs and interdisciplinary perspectives received in the course of the DeSeCo project, we have constructed a threefold categorisation of key competences for interpreting and further conceptualising key competences. The three theory-grounded, broad categories are acting autonomously, using tools interactively and interacting in socially heterogeneous groups.

Within the three categories we have highlighted a number of key competences that are particularly relevant for coping with many of the complex demands individuals and societies face. The identified competences are the result of an analysis of the various lists of key competences presented in the country reports and the expert contributions in light of the established definitional, conceptual and normative criteria described above. In the terminology of human capital, they contribute to important outcomes such as economic resources/employment, political participation, economic growth, social cohesion and the realisation of human rights.

The three categories and the key competences identified within each category, briefly described below, are detailed in DeSeCo's final report (Rychen and Salganik, 2003).

4.2.1. Acting autonomously

To act autonomously incorporates two central interrelated ideas, the development of personal identity and the exercise of relative autonomy in the sense of deciding, choosing and acting in different social fields. Acting autonomously does not mean functioning in social isolation. Instead,

it implies that individuals are enabled to manage their lives in meaningful ways by exercising control over their living and working conditions and by playing an active part in shaping their own lives. It requires an orientation towards the future and an understanding of one's environment, its functioning and the roles one plays within that environment (Perrenoud, 2001). The following key competences are relevant when it comes to acting autonomously:

- (a) the ability to defend and assert one's rights, interests, limits and needs;
- (b) the ability to form and conduct life plans and personal projects;
- (c) the ability to act within the big picture/the larger context (i.e. the larger historical, cultural or environmental context of actions and decisions and their long-term and indirect consequences).

4.2.2. Using tools interactively

The word 'tool' is used in the broadest sense of the term to include language, information and knowledge in addition to physical tools. To use a tool interactively assumes not only a familiarity with the tool itself but also an understanding of how the tool changes the way one can interact with the world. A tool is not just a passive mediator but is instrumental as part of an active dialogue between the individual and his or her environment (Haste, 2001). Underlying this is the idea that we encounter our world through our cognitive, social and physical tools. These encounters shape how we make sense of the world and become competent in interaction. The following key competences have been identified as relevant when it comes to using tools interactively:

- (a) the ability to use language, symbols and text interactively;
- (b) the ability to use knowledge and information interactively (to manage and use it as a basis for understanding options, forming opinions, making decisions and for taking informed actions);
- (c) the ability to use (new) technology interactively (including the ability to see potential new uses of it).

4.2.3. Interacting in socially heterogeneous groups

In this category the focus is on the interaction with the 'other', the different other. Human

beings are dependent on ties to others for material and psychological survival, for a sense of self, identity and social meaning (Ridgeway, 2001). Living and participating in multicultural societies and coping with increasing individual and social diversity require the ability to interact and coexist with people who do not necessarily speak the same language (literally or metaphorically) or share the same memory or history. By bridging these divides, these interpersonal or social competences help create social capital. The following key competences are relevant when it comes to interacting effectively with other people:

- (a) the ability to relate well to others (involving, initiating and maintaining relationships with family members, friends, neighbour, boss, colleague, etc.);
- (b) the ability to cooperate (to work together toward a common goal);
- (c) the ability to manage and resolve conflict (in a constructive manner).

4.3. Interacting of key competences

The hypothesis is that for individuals to meet the complex demands of modern life in a successful way and to contribute to the quality of society, they need to mobilise these kinds of competences in particular contexts. The competences do not, however, operate independently of each other. They are interrelated. The resulting patterns of these interrelated competences can be conceived as constellations, whose forms or interplay – the particular relevance of each competence within this constellation – depend on the desired outcomes at the individual and societal level and the context in which they apply.

One way to illustrate the relative importance of key competences with regard to their contribution to desired outcomes in different contexts is by conceptualising a multi-dimensional space whose axis are defined by the key competences. The various contexts (a country or a social field) can then be located in the space depending on the relative importance of the key competences for meeting the demands or goals in question.

5. Implications for a coherent assessment strategy

Developing assessments and indicators for key competences is not in DeSeCo's scope of work, although such work is clearly implicated by the notions discussed above. At this point, with the establishment of a conceptual frame of reference for key competences, it is possible and appropriate to highlight some of those implications for the development of assessments and indicators. This section presents some issues that are relevant to the elaboration of a long-term data strategy for evaluating learning outcomes in the OECD.

5.1. Dealing with the complex nature of key competences

The three broad categories of key competences, acting autonomously, using tools interactively and interacting in socially heterogeneous groups, have been recognised in the OECD as a valuable conceptual tool for mapping and further conceptualising the competences for which statistical indicators are developed. DeSeCo's work also confirms the theoretical relevance of what has been measured or proposed to be measured in international comparative assessment projects such as IALS, PISA and ALL. With the exception of initial efforts related to acting autonomously (self-regulated learning in PISA) and interacting in socially heterogeneous groups (exploration of teamwork for ALL), these projects have focused on competences which are related to using tools interactively (mainly related to using language, symbols and texts). This is an area for which theory provides a strong basis for measurement. Future research efforts in the field of key competence should be directed to the development and validation of measures related to key competences identified in the other two categories, acting autonomously and interacting in socially heterogeneous groups. However, key competences in these areas are much more socially and culturally embedded, a fact that must be taken into account in the development of any measures

and particularly internationally comparable measures (Murray, 2003). Focusing on new competence domains does not, of course, preclude continued research and development of the competences included in the category using tools interactively. In particular, research is required with respect to representing the level of mental complexity that involves critical thinking and a reflective, holistic and integrated approach.

In all categories of key competences, developing appropriate and reliable measures and interpreting the significance of any results necessitates open and innovative approaches to assessment methodologies, for several reasons. One is that assessments ideally reflect the complex nature of a competence, a combination of different interrelated component elements, including knowledge, cognitive skills and social and behavioural components (e.g. motivation and ethical/moral attitudes). This requires either some method of assessing the competence itself, or assessing one or more of the component elements and developing a model to relate the results to statements about the level of competence development.

Second, because key competences are inter-related and operate as constellations that vary with cultural and contextual factors, assessing key competences in isolation will not do justice to the multi-dimensional reference points associated with key competences. A variety of methodologies and settings will need to be explored to more accurately capture the complex nature of a competence and effectively profile the constellations of key competences that contribute to highly valued outcomes.

Third, similar to the more recent conceptualisations of 'literacy' and unlike traditional notions of 'literate' and 'illiterate', assessing competence is not a matter of assessing whether an individual does or does not possess it, but rather of determining where along a continuum of competence an individual's performance falls. Such a continuum must include a comprehensive range of levels to reflect the implications of the necessary level of mental complexity, in which individ-

uals ground their decisions and actions in critical thinking and a reflective/integrative/holistic approach.

Hence, the findings from DeSeCo reinforce the importance of further developing complementary assessment methodologies to produce a system for providing policy-relevant information on key competences and to more accurately capture the expression of key competences. The complex nature of key competences – including such aspects as their demand orientation, interrelatedness and relation to broadly desired outcomes – requires data from multiple sources, including, but not limited to, the collection of large-scale assessment data through paper and pencil or computerised instruments. Interactive computer simulations hold particular promise, as they allow for broadening assessment tasks and creating dynamic assessment situations. Adaptive and interactive testing simulates a more authentic context, where an individual's actions may be judged not only in response to a single static prompt, but progressively, on multiple levels as a situation develops. In addition, performance assessments and portfolios provide other complementary approaches for reviewing individuals' development and progress in one or more areas over time. And, last but not least, video-based data collection, as pioneered in TIMSS could provide more in-depth and qualitative information and add to a more complete picture of competences within various populations.

It should be noted that assessments using these methods are not well established and are almost certainly more expensive than traditional assessments. This should not be viewed as an insurmountable barrier to developing indicators for key competences though, but rather as an indication of the challenges faced.

5.2. Towards a coherent assessment strategy

Good governance and effective policy-making rely on information, data and statistics that are based on sound foundations. The design of a

coherent international assessment strategy for competences therefore needs to be guided by policy priorities and by an overarching conceptual frame of reference such as that provided by DeSeCo. With specific regard to key competences, policy-makers need to reach consensus on the value of choosing to measure some key competences or components of competences over others (considering factors such as the extent to which they are amenable to policy intervention). At the same time, researchers need to provide further conceptual input during that selection process and effective guidance on the feasibility of developing valid and reliable measures of new theory-grounded constructs. A future coherent assessment programme should take advantage of the conceptual and empirical frameworks that currently support internationally valid, reliable and comparable assessments of competences, building on the empirical experiences of current large-scale studies, in particular PISA and ALL.

A forward-looking assessment strategy should include plans and methods not only for providing data on the level and distribution of key competences over time and in different contexts, but also for using that information to learn more about the interrelatedness of key competences and their influence on desired outcomes. Containing this type of information, competence-related assessments could thus be used to answer such questions as, 'Do certain sub-populations show higher levels of development in certain key categories?' 'How has the distribution of key competences changed over time and what are possible explanations for these changes?' 'Are there correlations between levels of development of certain key competences?' and 'Do key competences in fact appear to lead to desired societal outcomes?'. Answering such questions would require a mechanism for collecting data on social contexts, a longitudinal approach, assessment of people from different backgrounds and at different stages in life and a cyclical structure of assessments. A coherent assessment programme, in turn, calls for coordinated planning and implementation at the international level and sustained political and financial support over many years.

6. Conclusions

The OECD project DeSeCo concluded with its final report entitled *Key competencies for a successful life and a well-functioning society* (Rychen and Salganik, 2003).

Through its collaborative process, DeSeCo succeeded in its missions: to propose a framework for understanding key competences based both on current theory from a variety of academic disciplines and on practical considerations. This overarching frame of reference serves several purposes: it presents a basis for a common, coherent and international discourse on competence development in a lifelong learning perspective; it has the potential to focus and improve future initiatives related to key competences, most significantly the development of a comprehensive strategy for data collection and analysis, including a determination of which competences to assess, both in the mid and long term; and to guide the development of assessment instruments and policy-relevant indicators.

Any such strategy must be linked to work already underway regarding domain-specific assessments. DeSeCo can help situate these assessments – and other empirical findings on

the outcomes of learning and teaching – in a broader conceptual context and clarify what is actually being measured and what is not.

Beyond assessments and indicators, the results of DeSeCo can provide valuable input to the prioritisation of particular competences for education curricula and training programmes development and for discussions of policies and programmes aimed at enhancing key competences among specific or general populations.

Although DeSeCo has provided some answers, it has raised new questions and provided a glimpse of the significant challenges that lie ahead. It is clear, for example, that further discussion and consensus-building around the recommendations of the project presented to the OECD are necessary (OECD, 2000). In particular, OECD Member countries need to establish broad consensus on new competence areas for assessment at the international level and start mapping out a coherent strategy for the assessment of key competences of youth and adults. A significant research effort is needed to make substantial strides in the breadth of new competences that can be meaningfully included in large-scale assessments.

List of abbreviations

ALL	Adult literacy and life skills
DeSeCo	Definition and selection of competences: theoretical and conceptual foundations
IALS	International adult literacy survey
INES	International indicators of education systems [OECD's long-term programme on education indicators]
PISA	Programme for international student assessment

Bibliography

- Binkley, M. et al. An overarching framework for understanding and assessing lifeskills. In: *Frameworks: working drafts*. Briefing materials for ALL Survey National Study Managers meeting held in Luxembourg, 23-24 September 1999.
- Callieri, C. The knowledge economy: a business perspective. In: Rychen, D. S.; Salganik, L. H. (eds) *Defining and selecting key competencies: theoretical and conceptual foundations*. Göttingen: Hogrefe and Huber, 2001, p. 228-231.
- Canto-Sperber, M; Dupuy, J.P. Competencies for the good life and the good society. In: Rychen, D. S.; Salganik, L. H. (eds) *Defining and selecting key competencies: theoretical and conceptual foundations*. Göttingen: Hogrefe and Huber, 2001, p. 67-92.
- Carson, J. Defining and selecting competencies: historical reflections on the case of IQ. In: Rychen, D. S.; Salganik, L. H. (eds) *Defining and selecting key competencies: theoretical and conceptual foundations*. Göttingen: Hogrefe and Huber, 2001, p. 32-44.
- Delors, J.; Draxler, A. From unity of purpose to diversity of expression and needs: a perspective from Unesco. In: Rychen, D. S.; Salganik, L. H. (eds) *Defining and selecting key competencies: theoretical and conceptual foundations*. Göttingen: Hogrefe and Huber, 2001, p. 214-221.
- DeSeCo. Available from Internet: www.deseco.admin.ch. [cited 30.8.2002]
- Farrugia, J.P. Competency management as an investment: a business perspective. In: Rychen, D. S.; Salganik, L. H. (eds) *Defining and selecting key competencies: theoretical and conceptual foundations*. Göttingen: Hogrefe and Huber, 2001, p. 93-120.
- Fratczak-Rudnicka, B.; Torney-Purta, J. Competencies for civic and political life in a democracy. In: Rychen, D. S.; Salganik, L. H.; McLaughlin, M. E. (eds) *Contribution to the second DeSeCo's symposium*. Neuchâtel, Switzerland: SFSO – Swiss Federal Statistical Office, 2003.
- Gilomen, H.; Rychen, D. S.; Salganik, L. H. Concluding remarks. In: Rychen, D. S.; Salganik, L. H. (eds) *Defining and selecting key competencies: theoretical and conceptual foundations*. Göttingen: Hogrefe and Huber, 2001, p. 247-251.
- Gonczi, A. Teaching and learning of the key competencies. In: Rychen, D. S.; Salganik, L. H.; McLaughlin, M. E. (eds) *Contribution to the second DeSeCo's symposium*. Neuchâtel, Switzerland: SFSO – Swiss Federal Statistical Office, 2003.
- Goody, J. Competencies and education: contextual diversity. In: Rychen, D. S.; Salganik, L. H. (eds) *Defining and selecting key competencies: theoretical and conceptual foundations*. Göttingen: Hogrefe and Huber, 2001, p. 175-190.
- Harris, B. Are all key competencies measurable? An education perspective. In: Rychen, D. S.; Salganik, L. H. (eds) *Defining and selecting key competencies: theoretical and conceptual foundations*. Göttingen: Hogrefe and Huber, 2001 p. 222-227.
- Haste, H. Ambiguity, autonomy and agency: psychological challenges to new competence. In: Rychen, D. S.; Salganik, L. H. (eds) *Defining and selecting key competencies: theoretical and conceptual foundations*. Göttingen: Hogrefe and Huber, 2001, p. 93-120.
- Keating, D. P. Definition and selection of competencies from a human development perspective. In: Rychen, D. S.; Salganik, L. H.; McLaughlin, M. E. (eds) *Contribution to the second DeSeCo's symposium*. Neuchâtel, Switzerland: SFSO – Swiss Federal Statistical Office, 2003.
- Kegan, R. Competencies as working epistemologies: ways we want adults to know. In: Rychen, D. S.; Salganik, L. H. (eds) *Defining and selecting key competencies: theoretical and conceptual foundations*. Göttingen: Hogrefe and Huber, 2001, p. 192-204.
- Levy, F.; Murnane, R. J. Key competencies critical to economic success. In: Rychen, D. S.; Salganik, L. H. (eds) *Defining and selecting key competencies: theoretical and conceptual*

- foundations*. Göttingen: Hogrefe and Huber, 2001, p. 151–174.
- Making a European area of lifelong learning a reality / European Commission. Luxembourg: Office for Official Publications of the European Communities, 2001 (COM(2001) 678 final).
- Murray, T. S. Reflections on future international assessments. In: Rychen, D. S.; Salganik, L. H. (eds) *Key competencies for a successful life and a well-functioning society*. Göttingen: Hogrefe and Huber, 2003, p. 135–160.
- Oates, T. Key skills/key competencies: avoiding the pitfalls of current initiatives. In: Rychen, D. S.; Salganik, L. H.; McLaughlin, M. E. (eds) *Contribution to the second DeSeCo's symposium*. Neuchâtel, Switzerland: SFSO – Swiss Federal Statistical Office, 2003.
- OECD. *Definition and selection of competencies (DeSeCo): theoretical and conceptual foundations* (Strategy Paper). Paris: OECD, 2002. Available from Internet: http://www.statistik.admin.ch/stat_ch/ber15/deseco/deseco_strategy_paper_final.pdf. [cited 6.6.2003]
- OECD. *The well-being of nations – the role of human and social capital*. Paris: OECD, 2001a.
- OECD. *Investing in competencies for all*. Communiqué for the meeting of the OECD Education Ministers, Paris, 3–4 April 2001. Paris: OECD, 2001b. Available from Internet: <http://www.oecd.org/pdf/M00008000/M00008906.pdf> [cited 6.6.2003]
- OECD. *Measuring student knowledge and skills: a new framework for assessment*. Paris: OECD, 1999.
- Ouane, A. Defining and selecting key competencies in lifelong learning. In: Rychen, D. S.; Salganik, L. H.; McLaughlin, M. E. (eds) *Contribution to the second DeSeCo's symposium*. Neuchâtel, Switzerland: SFSO – Swiss Federal Statistical Office, 2003.
- Perrenoud, P. The key to social fields: competencies of an autonomous actor. In: Rychen, D. S.; Salganik, L. H. (eds) *Defining and selecting key competencies: theoretical and conceptual foundations*. Göttingen: Hogrefe and Huber, 2001, p. 121–150.
- Ridgeway, C. Joining and functioning in groups, self-concept and emotion management. In: Rychen, D. S.; Salganik, L. H. (eds) *Defining and selecting key competencies: theoretical and conceptual foundations*. Göttingen: Hogrefe and Huber, 2001, p. 205–211.
- Ritchie, L. Key Competencies for whom? A labor perspective. In: Rychen, D. S.; Salganik, L. H. (eds) *Defining and selecting key competencies: theoretical and conceptual foundations*. Göttingen: Hogrefe and Huber, 2001, p. 236–240.
- Rychen, D. S. Introduction. In: Rychen, D. S.; Salganik, L. H. (eds) *Defining and selecting key competencies: theoretical and conceptual foundations*. Göttingen: Hogrefe and Huber, 2001, p. 1–15.
- Rychen, D. S.; Salganik, L. H. (eds) *Defining and selecting key competencies: theoretical and conceptual foundations*. Göttingen: Hogrefe and Huber, 2001.
- Rychen, D. S.; Salganik, L. H. (eds) *Key competencies for a successful life and a well-functioning society*. Göttingen, Germany. Hogrefe and Huber, 2003.
- Rychen, D. S.; Salganik, L. H.; McLaughlin, M. *Contribution to the second DeSeCo's symposium*. Neuchâtel, Switzerland: SFSO – Swiss Federal Statistical Office, 2003.
- Salganik, L. H. Competencies for life: a conceptual and empirical challenge. In: Rychen, D. S.; Salganik, L. H. (eds) *Defining and selecting key competencies: theoretical and conceptual foundations*. Göttingen: Hogrefe and Huber, 2001, p. 17–32.
- Salganik, L. H. et al. *Projects on competencies in the OECD context: analysis of theoretical and conceptual foundations*. Neuchâtel, Switzerland: SFSO – Swiss Federal Statistical Office, 1999.
- Stein, S. *Equipped for the future content standards: what adults need to know and be able to do in the 21st century*. Washington, DC: National Institute for Literacy, 2000.
- SFSO – Swiss Federal Statistical Office. *Country contribution process: summary and country reports*. Briefing materials prepared for DeSeCo's 2nd International Symposium, Geneva, 11–12 February 2002. Neuchâtel, Switzerland: SFSO, 2001a. (CCP Reports) Available from Internet: http://www.statistik.admin.ch/stat_ch/ber15/deseco/deseco_country.htm [cited 6.6.2003]
- SFSO – Swiss Federal Statistical Office. *DeSeCo background paper*. Neuchâtel, Switzerland: SFSO, 2001c. Available from Internet:

http://www.statistik.admin.ch/stat_ch/ber15/deseco/deseco-background-paper02.pdf [cited 6.6.2003]

The World Bank. *Lifelong learning in the global knowledge economy: challenges for developing countries*. 2002. Available from Internet: http://www1.worldbank.org/education/pdf/Lifelong%20Learning_GKE.pdf [cited 6.6.2003]

Trier, U. P. Defining educational goals: a window to the future. In: Rychen, D. S.; Salganik, L. H. (eds) *Defining and selecting key competencies: theo-*

retical and conceptual foundations. Göttingen: Hogrefe and Huber, 2001, p. 241-246.

Trier, U. P. Twelve countries contributing to DeSeCo: a summary report. In: Rychen, D. S.; Salganik, L. H.; McLaughlin, M. E. (eds) *Contribution to the second DeSeCo's symposium*. Neuchâtel, Switzerland: SFSO – Swiss Federal Statistical Office, 2003.

Weinert, F. E. Concept of competence: a conceptual clarification. In: Rychen, D. S.; Salganik, L. H. (eds) *Defining and selecting key competencies: theoretical and conceptual foundations*. Göttingen: Hogrefe and Huber, 2001, p. 45-66.

International symposia proceedings and papers

DeSeCo Symposium, Neuchâtel, October 13-15, 1999. Available from Internet:

http://www.statistik.admin.ch/stat_ch/ber15/deseeco/deseeco_symp99.htm

DeSeCo Symposium, Geneva, February 11-13, 2002. Available from Internet:

http://www.statistik.admin.ch/stat_ch/ber15/deseeco/deseeco_int02.htm

Further information

DeSeCo Secretariat
Dominique Simone Rychen
Program Manager
Swiss Federal Statistical Office
Espace de l'Europe 10
CH-2010 Neuchâtel

Phone: +41 (0)32 713 61 60
Email: dominique.rychen@bfs.admin.ch
Internet: www.deseeco.admin.ch

List of authors

Elliot Stern

The Tavistock Institute
UK – London EC2A 4UE
E-mail: e.stern@tavinstitute.org

Wolfgang Beywl**Sandra Speer**

Univation Evaluation Institute
D – 50674 Köln
E-mail: wolfgang.beywl@univation.org

Reinhard Hujer**Marco Caliendo****Dubravko Radić**

Johann Wolfgang Goethe University Frankfurt
am Main
D – 60325 Frankfurt am Main
E-mail: hujer@wiwi.uni-frankfurt.de

Evelyn Viertel**Søren P. Nielsen****David L. Parkes****Søren Poulsen**

European Training Foundation
I – 10133 Torino
E-mail: Evelyn.Viertel@etf.eu.int

Beatriz Pont**Patrick Werquin**

OECD
F – 75775 Paris Cedex 16
E-mail: Patrick.WERQUIN@oecd.org

Gerald A. Straka

University of Bremen
D – 28359 Bremen
E-mail: straka@uni-bremen.de

Dominique Simone Rychen

Swiss Federal Statistical Office, SFSO
CH – 2010 Neuchâtel
E-mail: dominique.rychen@bfs.admin.ch

Cedefop (European Centre for the Development of Vocational Training)

The foundations of evaluation and impact research
Third report on vocational training research in Europe: background report

Pascaline Descy, Manfred Tessaring (eds)

Luxembourg: Office for Official Publications of the European Communities

2004 – 332 pp. – 21.0 x 29.7 cm

(Cedefop Reference series; 58 – ISSN 1608-7089)

ISBN 92-896-0278-3

Price (excluding VAT) in Luxembourg: EUR 25

No of publication: 3040 EN

Renowned researchers have contributed to this publication covering philosophical roots, types and standards of evaluation and impact research. It also discusses tools and methods for evaluating education and training systems and reforms are discussed. Competence measurement as well as design and selection of key competences in an international context are also addressed.

This publication is part of the background report to Cedefop's third research report, dedicated to evaluation and impact of education and training. The other volumes address the *Impact of education and training* and the *Evaluation of systems and programmes*. The third research report is complemented by a synthesis report which provides a comprehensive review of types and results of evaluation and impact research in Europe and beyond.

The series of Cedefop reports on vocational education and training research have been published since 1998. They take into account the wider context in which vocational education and training is embedded and discuss the implications for policy, practice and future research.

Pascaline Descy, Manfred Tessaring

EN

The foundations of evaluation and impact research

**Third report on vocational
training research in Europe**
Background report



European Centre for the
Development of Vocational Training

Europe 123, GR-570 01 Thessaloniki (Pylea)
Postal address: PO Box 22427, GR-551 02 Thessaloniki
Tel. (30) 23 10 49 01 11, Fax (30) 23 10 49 00 20
E-mail: info@cedefop.eu.int
Homepage: www.cedefop.eu.int
Interactive website: www.trainingvillage.gr

Price (excluding VAT) in Luxembourg: EUR 25

3040 EN



Publications Office
Publications.eu.int

ISBN 92-896-0278-3



9 789289 602785 >