Developing standards to evaluate vocational education and training programmes

Wolfgang Beywl; Sandra Speer

In:

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

The foundations of evaluation and impact research

Third report on vocational training research in Europe: background report.
Luxembourg: Office for Official Publications of the European Communities, 2004
(Cedefop Reference series, 3040)

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

For your information:

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

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

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

For further information contact:

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

Contributions to the background report of the third research report

Impact of education and training

Preface

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

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

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

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

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

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

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

The foundations of evaluation and impact research

Preface

Philosophies and types of evaluation research
Elliot Stern

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

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

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

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

Measurement and evaluation of competence
Gerald A. Straka

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

Evaluation of systems and programmes

Preface

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

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

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

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

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

Evaluation activities in the European Commission
Josep Molsosa

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

List of tables and figures

Tables

Figures

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

([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.

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. Nuissl (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 European 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 evaluand 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?

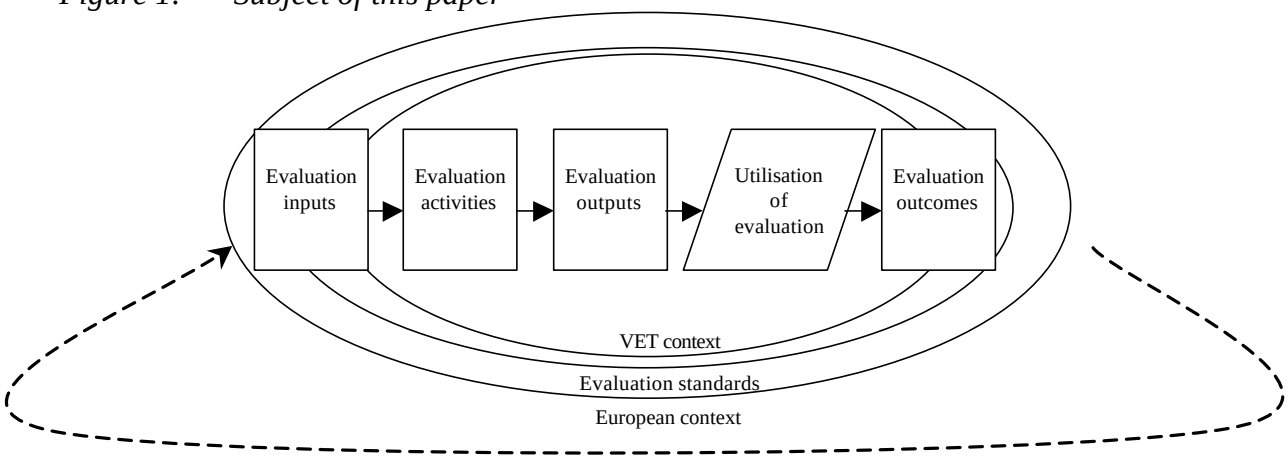
(f) are there quality requirements for evaluations in VET contexts which are missing in general evaluation standards? Should there be additional standards or extensions of existing standards?

(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.

Figure 1: Subject of this paper



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

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.

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.

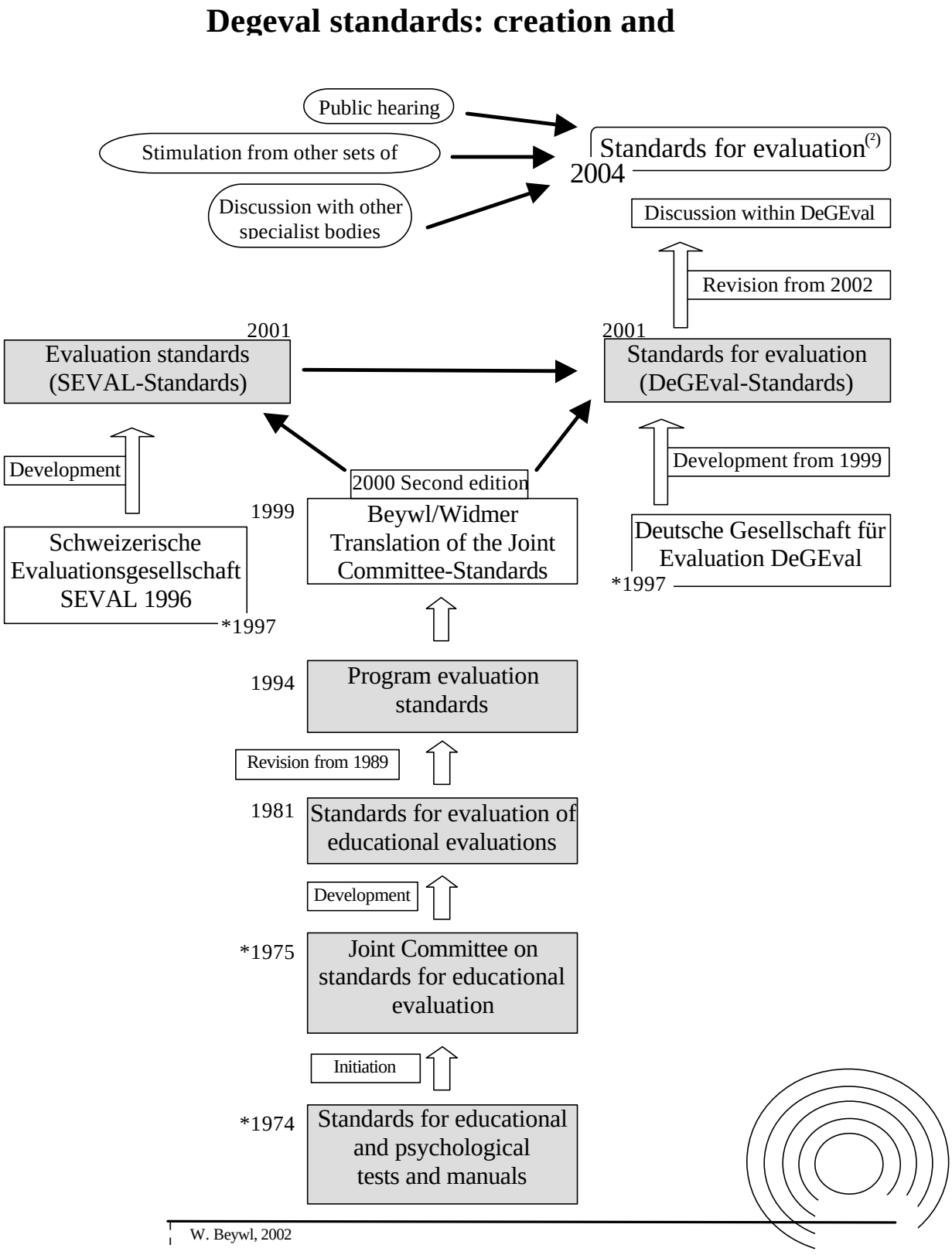
2.1. Background and purpose of evaluation standards

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’.

Figure 2: Evolution of DeGEval standards



Source: DeGEval (2002, p.2; slightly revised by the author)

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]).

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

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

([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.

(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.

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 practicians 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

([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.

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).

([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).

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 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 standard' 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,

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

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 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

---

([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*.

([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-achieveability 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.

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

([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.

evaluators' professional and personal skills and their adherence to general laws and codes of ethics and assumption of professional and personal responsibility ([16]).

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 acknowledgement 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.

([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).

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

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 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

([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).

([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.

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.

([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).

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

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

([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).

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. Evaluations 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).

([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.

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

---

([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.

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 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 1 645 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);

([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.

([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).

(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]).

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

([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).

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 timeliness' 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

([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.

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 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

([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].

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 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

---

([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. 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 evaluation 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, 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.

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

([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).

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 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.

([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.

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).

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

(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

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

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 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, fundaments 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 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).

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

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])?

([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.

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 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, transdisciplinary 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

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

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

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 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

([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.

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?

([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].

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;

(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

([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.

([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.

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 evaluand 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 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

([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.

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 situation 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]).

([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 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 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/evaluand (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.

([65]) See the questionnaire in Annex 2

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

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.

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.

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

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

Figure 3: Respondents' identification with national professional cultures

Source: author's representation

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

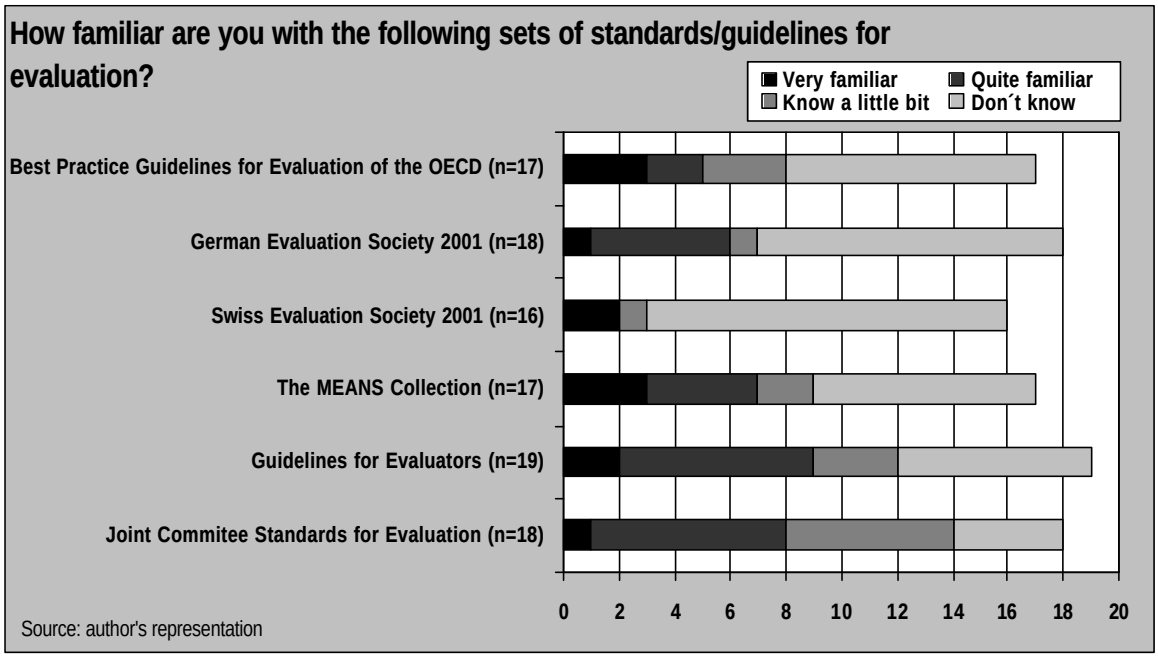
Over half (10 respondents) of the international experts said that they were familiar with VET but that they were rarely, if at all, involved in it. Around a third (six) identified VET as a relevant working field for them. Only three respondents named VET as their main or most relevant working field. 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



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 AFPD, IEFP 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]).

([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.

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;

(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.

([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.

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 loose 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;

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

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. 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.

([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.

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:

(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.

([74]) See Summary and Outlook.

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 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

([75]) See answers to question 6.

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.

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

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

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.

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.

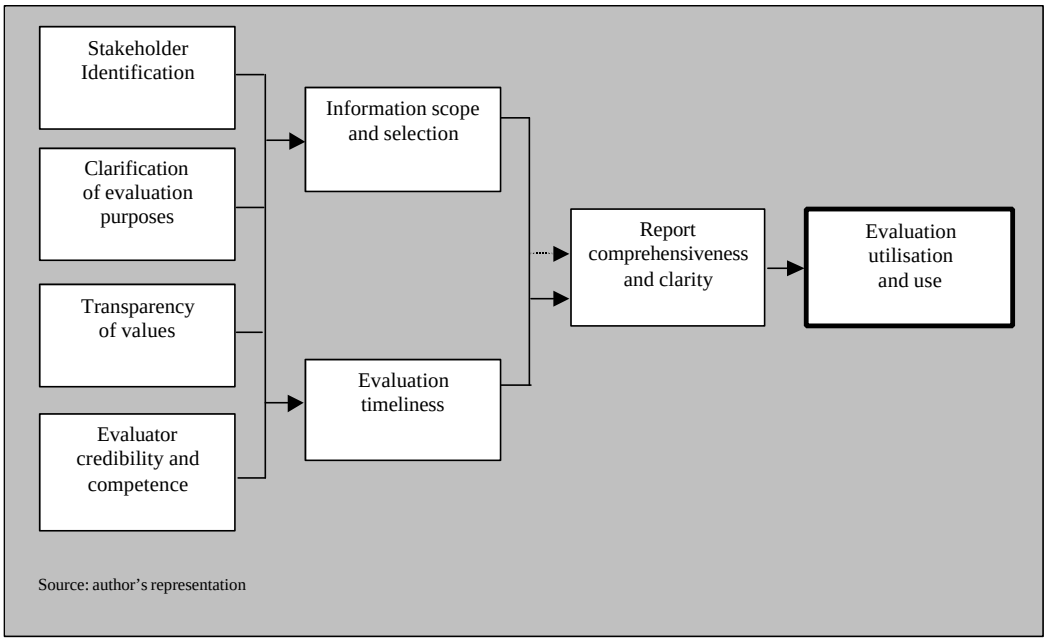
Below are the 25 DeGEval standards, grouped according to the four standard categories, with notes gleaned from European VET-oriented literature. We have refrained from providing a detailed description of each standard, as this appears in the attached version of the DeGEval standards and their explanatory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



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 evaluand 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]).

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);

([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.

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

(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 the evaluation team members. Objectivity and independence are the other key considerations for selecting evaluators. The degree of independence, however, depends on who

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

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

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.

'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

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

([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.

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 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).

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

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

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

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.

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.

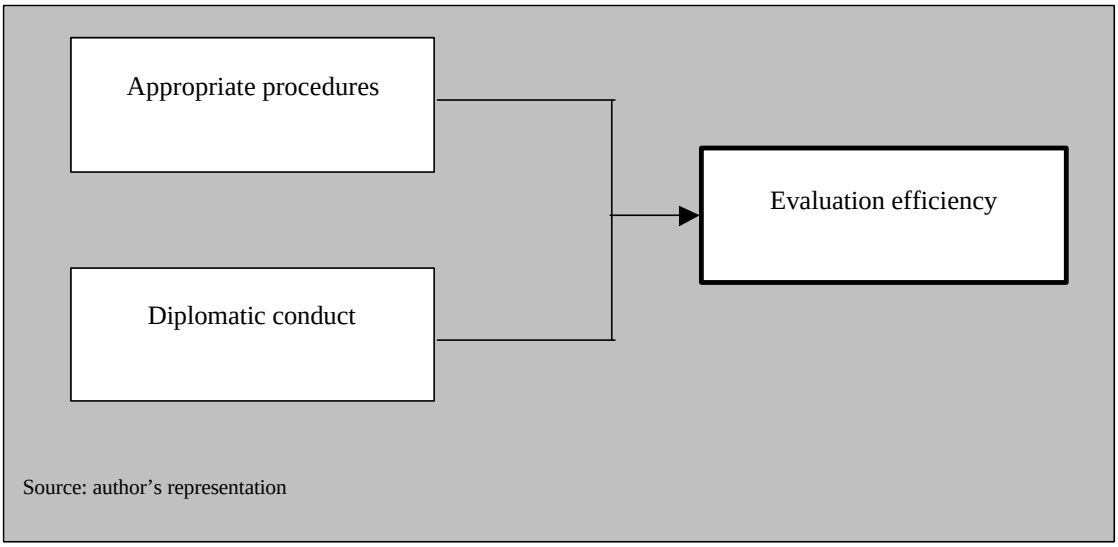
([85]) See Auer and Kruppe (1996) for an overview of monitoring systems in the EU.

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 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).

Figure 6: Two points which make evaluations feasible



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);

(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 implemented 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.

([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.

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 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?

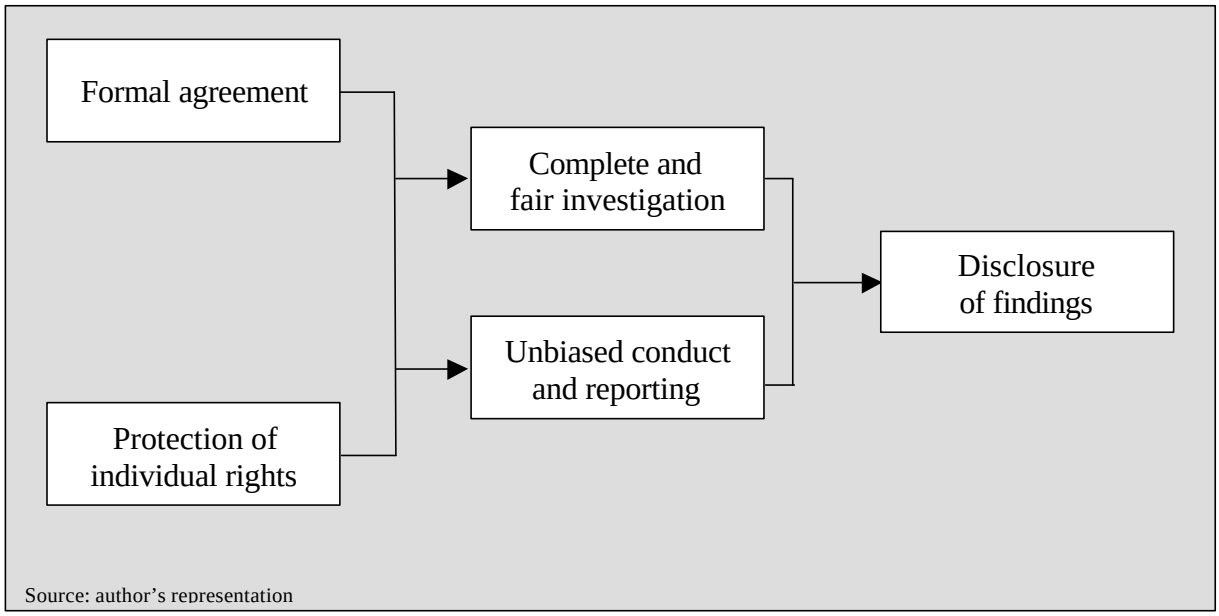
(g) is evaluation of the training or development measure (urgently) needed?

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

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



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.

([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 …).

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 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 partners 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]).

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

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 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.

([89]) The SEVAL K6 standard Declaration of conflicts of interests is formulated 'more realistically' than the DeGEval standard.

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).

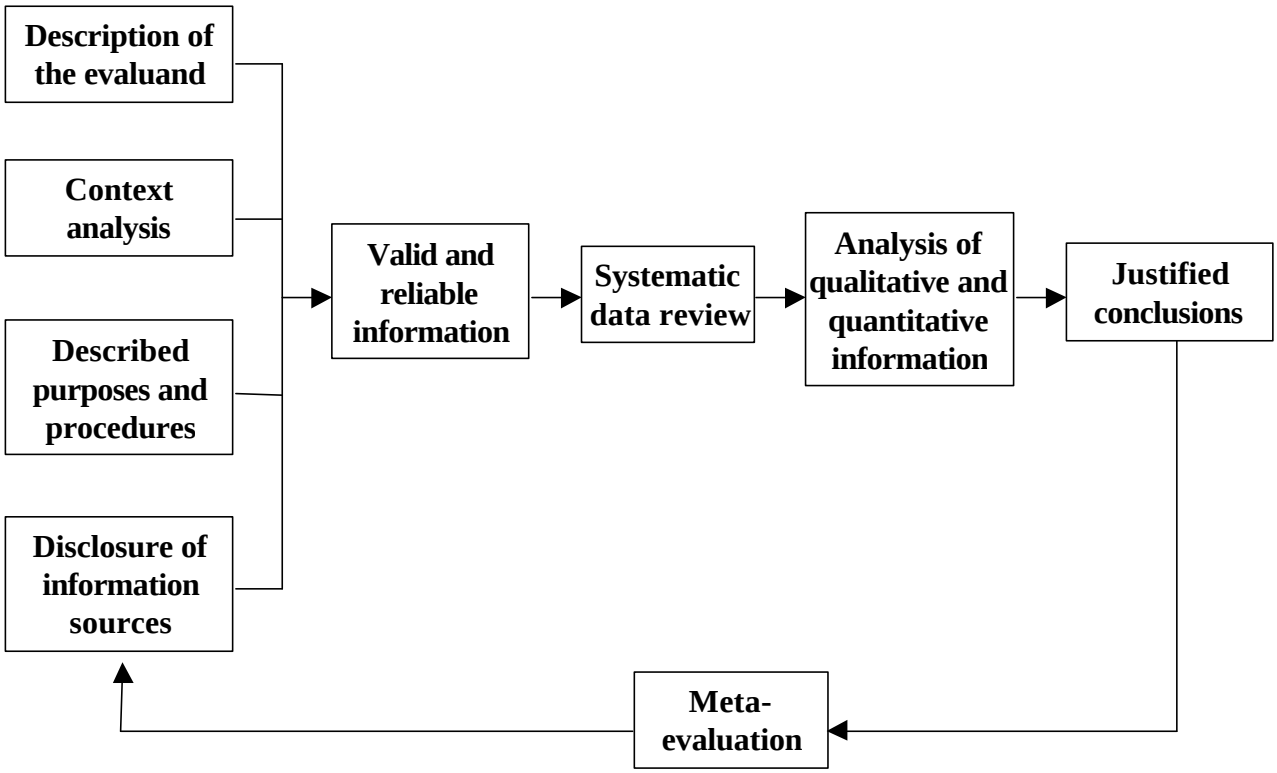
The nine standards in this group can be broken down into four categories. The first two standards (G1/A1 and G2/A2) address the definition of the evaluand in context and demand a description of it. The next two standards (G3/A3 and G4/A4) demand identification of purpose, procedures and the information 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]).

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

([91]) ibid.

Figure 8: Nine components which make evaluations accurate



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

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

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 considerable 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 evaluand 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.

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

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.

([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.

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 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. 540I). 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 evaluand 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	*Deutsche Gesellschaft für Evaluation* [German Evaluation Society]
EES	European Evaluation Society
ISO	International Organisation for Standardisation
JC	Joint Committee on Standards for Educational Evaluation
SEVAL	Schweizerische Evaluationsgesellschaft [Swiss Evaluation Society]
SFE	Société Française de l'Évaluation [French Evaluation Society]

Annex 1: transformation table

Deutsche Gesellschaft für Evaluation		**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

Univation e.V.
Institute for Evaluation
http://www.univation.org

Zuelpicher Strasse 58
D-50674 Koeln
Germany

Tel: +49(0)221/4248071
Fax: +49(0)221/4248072
E-mail: cedefop@univation.org

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).

Univation e.V.
Institute for Evaluation
http://www.univation.org

Zuelpicher Strasse 58
D-50674 Koeln
Germany

Tel: +49(0)221/4248071
Fax: +49(0)221/4248072
E-mail: cedefop@univation.org

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: .....……

4) What is your relation to vocational educational and training (VET)? *(choose one alternative)*

   ☐ VET is my main/most relevant working field

   ☐ VET is one of my most relevant working fields

   ☐ VET is a known field for me but I am (nearly) not active in VET

5) What are, in the nearer and distant future, the strongest competitors of evaluation in VET in the country you mainly work in (if you work on an international level please answer the question for the EU and its Member States)?

   (Please mark the best fitting category in each row)

Univation e.V.
Institute for Evaluation
http://www.univation.org
Zuelpicher Strasse 58
D-50674 Koeln
Germany
Tel: +49(0)221/4248071
Fax: +49(0)221/4248072
E-mail: cedefop@univation.org

	Very strong	**Strong**	**Weak**	**Very weak/ not existing**
Auditing	☐	☐	☐	☐
Benchmarking	☐	☐	☐	☐
Certification/accreditation	☐	☐	☐	☐
Monitoring	☐	☐	☐	☐
Performance/results based management	☐	☐	☐	☐
Quality management/assurance	☐	☐	☐	☐
State supervision	☐	☐	☐	☐
Other: ……..	☐	☐	☐	☐
Other: ……..	☐	☐	☐	☐

6) If you would describe your general position to **s**tandards for evaluation, which of the following statements would mostly express your opinion? *(Choose one alternative)*

☐ Standards for evaluation are not necessary or even detrimental

☐ Standards for evaluation do not matter

☐ Standards for evaluation could be useful; but I am not convinced whether they will be in fact

☐ Standards for evaluation are important

☐ Standards for evaluation are absolutely necessary

7) Preferred type of standards *(If you answered 'unnecessary' or 'do not matter' in question 6, skip this question)*

There are two distinct concepts of standards

Minimum standards (as in engineering or work security): they describe 'features of evaluation in a very precise, operational way; if one ore more standards are not fulfilled, the evaluation will be judged as 'poor' or 'non-professional'.

Maximum standards (as in education or consulting) are standards one should strive for; it should be clearly justified if one or more standards are not taken into account in evaluation practice. Some standards not considered within an evaluation would not automatically lead to a negative judgement of the evaluation as a whole.

Univation e.V.
Institute for Evaluation
http://www.univation.org
Zuelpicher Strasse 58
D-50674 Koeln
Germany
Tel: +49(0)221/4248071
Fax: +49(0)221/4248072
E-mail: cedefop@univation.org

Which kind of standards do you prefer for evaluation? *(Choose one alternative)*

	Strongly prefer	**Prefer**	**Cannot decide**	**Prefer**	**Strongly prefer**	
Minimum standards	☐	☐	☐	☐	☐	Maximum standards

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.

……
……

Univation e.V.
Institute for Evaluation
http://www.univation.org
Zuelpicher Strasse 58
D-50674 Koeln
Germany
Tel: +49(0)221/4248071
Fax: +49(0)221/4248072
E-mail: cedefop@univation.org

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	*Centre d'Études de l'Emploi* 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	*Sozialforschungsstelle Dortmund Landesinstitut* 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	*Centre d'études et de recherches sur les qualifications* 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	

Name, surname Function in VET/evaluation	Institution, city, country, e-mail	Other functions in VET/evaluation
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. Deontolgia de la evaluación: el modelo de los códigos éticos anglosajones. *Gestión y Análisis de Politica 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.finnishevaluationsociety.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.

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.

Gullickson A R. The student evaluation standards: how to improve evaluations of students. Thousand Oaks: Corwin Press, 2002.

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: Josses-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/GuidanceGoodPractice.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].