INSTITUTE OF TECHNOLOGY AND REGIONAL POLICY

INTEREG RESEARCH REPORT SERIES

InTeReg Research Report Projekt Nr. RTW.2003.AF.022-01

EU EVALUATION PRATICE AND TOOLKIT APPLICABILITY FOR JAPAN

Luke Georghiou [1)], Wolfgang Polt [2)]

[1)] PREST, University of Manchester, [2)] Joanneum Research

July 2004

JOANNEUM RESEARCH Forschungsgesellschaft mbH – Institut of Technology and Regional Policy (InTeReg)

Vienna Office:
Wiedner Hauptstraße 76
A-1040 Vienna, Austria
Phone: +43-1-581 75 20
E-Mail: intereg@joanneum.at

Graz Office:
Elisabethstraße 20
A-8010 Graz, Austria
Phone: +43-316-876 1488
E-Mail: intereg@joanneum.at

EU Evaluation Practice and Toolkit

Applicability for Japan

Study on behalf of

Mitsubishi Research Institute, Japan

EU Evaluation Practice and Toolkit – Applicability for Japan

1 Introduction

This paper aims to describe the practice and development of evaluation of research and development and related activities in the programmes of the European Commission and the methods available more generally before concluding by assessing the applicability of these approaches to the Japanese situation. Section 2 concentrates on evaluation practice in the Commission itself while Section 3 summarises the main elements of an evaluation toolbox developed on behalf of the European Commission but both drawing upon wider experiences and aiming to support evaluation in a similarly broad range of situations. Finally, Section 4 considers the relevance of the activity described in the preceding two sections for the practice of evaluation of similar activities in Japan.

2 Trends in Evaluation in the European Union

2.1 Framework for Evaluation in the European Commission

Establishment of the System

The origins of the European Commission's approach to evaluation of research and technological development programmes (RTD) go back more than two decades but still remain influential today. The emergence of the system is described in detail in Georghiou (1995)[1] and Guy and Arnold (1996)[2]. Before the 1980s a series of experimental activities had already established the main format which still continues in part today. This consists of a variant of the peer review approach. Independent experts, principally from the fields concerned but not participating in the programmes themselves, were constituted into Panels which met monthly over 6-8 months. Panels then reviewed the results of surveys and interviewed key actors. The terms of reference addressed three main criteria:

- Scientific and technological quality of research
- Effectiveness of programme management
- Contribution of results to the progress of science and technology

A centralised Evaluation Unit was established by DGXII in the early 1980s and has been in charge of the development of new and improved evaluation schemes and their implementation ever since. This unit developed the first multi-annual Plan of Action for evaluation in 1983, which initiated the build up to a fully operational evaluation system. It encouraged research on evaluation methodology through the SPEAR-MONITOR programme (e.g. research on the development of performance and impact indicators; bibliometric approaches; survey techniques etc.) and promoted the exchange of information on evaluation.

The Second Plan of Action (1987-1991) helped further institutionalise the evaluation process. Each Specific Programme within the overall Framework Programme was

[1] Georghiou, L. (1995), 'Assessing the Framework Programmes - a meta-evaluation', Evaluation, Vol. 1, Number 2, October 1995

[2] Guy K and Arnold E, Strategic Options for the Evaluation of the R&D Programmes of the European Union, Final Report EP/IV/B/STOA/96/A.LP.1 European Parliament STOA 1996

evaluated by external experts at least once during the period, the execution of the evaluation timed to influence activities scheduled for the next period. Panel formulation also became more heterogeneous, with membership drawn from experts in the technical fields concerned, experts from different fields, users of results, experts with management experience, science policy experts and evaluation specialists. Interim and final evaluations conducted by the Commission were also reported to the European Parliament and to the European Council.

Not all activities falling within the context of the Framework Programmes fell under the aegis of DGXII, however. DGXIII, for example, which administers programmes concerned with Information and Communication Technologies, evolved its own more decentralised approach to programme evaluation. The first evaluation to be conducted was a mid-term review of ESPRIT in 1985 (when it was the responsibility of the IT Task Force). A panel of experts supported by a small secretariat held face-to-face meetings with relevant organisations and conducted a questionnaire survey of participants. Other evaluations followed much the same pattern, i.e. programme audits carried out by groups of independent technical experts drawn mainly from the sectors concerned, with a focus on reviews of global objectives and priorities, studies of future requirements and options, and a technical audit of projects. Many of these evaluations have been criticised, however, for the limited attention they gave to issues of impact and appropriateness, and for the dominating presence on panels of major programme client groups (Georghiou, 1995).

A Legal Framework

In 1994, with the aim of meeting legislative requests in connection with the Fourth Framework Programme, the Commission introduced a new evaluation scheme based on continuous monitoring reporting annually and a five-year assessment carried out midway through programme implementation so that it included two previous programmes and produced results in time for the preparation of the next Framework Programme proposal. Thus each evaluation would be an ex post evaluation of the previous programme, a mid-term appraisal of the current programme and provide recommendations for future activities. This approach was endorsed in 1995 by CREST, a body of Member State representatives advising the Commission and council on S&T-related matters.

The system operates both at the level of sub-programmes and for the Framework Programme as a whole.

Continuous monitoring is carried out with the *assistance* of a panel of external experts while 5 year assessments are *conducted* by the experts. The Commission describes its Continuous monitoring procedures as set out in Box 1 below:

Box 1 Continuous Monitoring

For the Specific Programmes, this exercise aims at ensuring the cost effective implementation, examining progress in relation with the original objectives and whether these objectives, priorities and financial resources are still appropriate to changing circumstances. For the Framework Programme, the aim is to monitor overall progress as regards the major objectives, and to examine whether the objectives, priorities and financial resources are still appropriate in the overall context.

The day-to-day monitoring of the Specific Programmes' implementation is carried out by the programme management within the Commission Services. In addition, panels of external experts provide an independent view once a year (the process itself lasts over a few months) on the progress of implementation. The panels, one for each of the Specific Programmes (and one for the Framework Programme) give advice on key issues relating to programme development and thus help programme management and programme committees to identify and correct weaknesses. In the light of the results of the monitoring, the Commission may submit proposals to adapt or supplement the Specific Programmes or the Framework Programme.

The continuous monitoring exercises aim also at collecting data which are useful for the five-years assessments. Moreover, the sets of the annual monitoring reports provide, notably, the five-years assessment panels with information on the effectiveness of programme implementation.

Both at Specific Programme and Framework Programme levels, the following issues are required to be addressed by the panels:

- efficiency and transparency of the programme management (including calls for proposals, information to applicants, the assessment and selection process, contract negotiation and disbursement of funds), and internal Commission co-ordination;
- consistency of the selection of projects with the initial objectives and the work programme, and the extent to which selected projects or clusters of projects fulfil the wider policy objectives of the European Union (in particular in areas of relevance to the programme concerned);
- use of specific measures and support activities (e.g. to support SMEs, improve dissemination, etc.), and participation in the programme of firms and institutions from less favoured regions;
- appropriate follow-up of previous evaluation/monitoring recommendations;
- the progress and output of projects against the original targets set; and
- aspects of flexibility to respond to the needs of society in the light of changing circumstances.

The panels are also invited to produce recommendations for the future indicators to be used for monitoring as well as for the monitoring process itself.
The Framework Programme level exercise is mainly a synthesis of the Specific Programmes' monitoring (including core indicators, summarising progress and giving emphasis to the main issues which have emerged from the analysis. Nevertheless, the panel's report is more than simply the sum of the specific programme monitoring reports. Therefore, the following additional issues are considered:

- cases where the independent monitoring experts consider the results will have a significant impact, or where poor performance requires further examination;
- as appropriate, consideration of Community RTD objectives as well as synergies between Specific Programmes; and
- changes that may be needed to the balance of the Programmes or to the strategy for implementation, in the light of experience and changes in the wider environment.

Source: CORDIS Fifth Framework Programme Monitoring Evaluation and Assessment Activities

Box 2 sets out the objectives and procedures for 5-Year assessments:

Box 2 5-Year Assessments

The overall objective of this exercise is to provide input to policy formulation and decision-making, based on feedback from implementation. For the Specific Programmes, the five-year

assessment aims at evaluating the activities carried out within the fields covered by the programmes and their management during the five years preceding the assessment. In particular, the following are examined: the relevance of the initial objectives considering major new developments, the cost-effectiveness of programme implementation, and effectiveness in achieving the original objectives. The exercise also results in identification of major achievements and lessons learned from programme implementation, and provides recommendations for future activities. It is expected that these evaluations can be carried out efficiently using the data collected during the monitoring operations and the monitoring reports themselves.

The Framework Programme five-year assessment, which is based on the Specific Programme assessments and combines them at a higher level, goes beyond the evaluation of past and current activities and considers the next Framework Programme taking into account also the working documents of the Commission available at that time. Consequently, this five-year assessment combines an ex-post evaluation of the previous Programme, a mid-term evaluation of the on-going one and an ex-ante appraisal of future activities.

Both at Specific Programme and Framework Programme levels, the following key issues are required to be addressed by the panels:

- relevance, i.e. whether the initial objectives are still valid against new S&T developments and socio-economic conditions;
- efficiency, i.e. whether the objectives have been pursued in a cost effective manner; and
- effectiveness, i.e. whether the initial objectives have been achieved, or, for longer term strategy and objectives, if progress is sufficient. Moreover, whether the "European added-value" has been adhered to and the results have been disseminated/exploited.

At the Framework Programme level, in addition, the panel is requested to pay attention to the coherence between the Community and national S&T policies with a view to enhancing their mutual consistency, and to aspects of co-ordination with other international S&T policies or programmes.

Source: CORDIS Fifth Framework Programme Monitoring Evaluation and Assessment Activities

Broader Requirements for Evaluation

The basic principles and requirements for evaluation in the European Commission as a whole were set out in *Commission's Communication on Evaluation in 1996 (SEC(96)659final)* which stated that every Directorate-General (DG) should have a designated evaluation function and annual evaluation plan, that actions financed on an annual basis were to be evaluated at least once every six years, and that multi-annual programmes should be subject to mid-term and ex post evaluations.

More recently evaluation has been seen as an important instrument in the context of reform of the European Commission, as set out in the document *Focus on Results: Strengthening Evaluation of Commission Activities SEC (2000) 1051*. This provides the context for evaluation in the reform era with which R&D evaluation has to comply. It also gives a definition of evaluation, first used in the White Paper on Reform:

> *"judgement of interventions according to their results, impacts and the needs they aim to satisfy"*

Key features presented in the document are:

- DGs and Services to carry out measures to improve existing evaluation systems and practices covering regular evaluation, setting up a dedicated evaluation function with the expertise to plan and manage evaluation, promoting quality, ex ante evaluation, better integration into decision-making, support to strategic programming and planning function, programme for evaluation and better monitoring.
- DG Budget (responsible for finance) co-ordinates an Evaluation Network, consisting of representatives from each DG's evaluation function. The Strategic Planning and Programming Function of the Commission's Secretariat General leads the development of standards for policy evaluation. Together they follow progress on implementation of the above features and promote good practice and exchange of experience.
- Quality of evaluation systems is assessed by DG Audit.
- Working group on Activity Based Management to assess results achieved in implementation of principles and define further measures if necessary.

2.2 Application in the Context of the F*ramework Programme*

Returning to the specific area of research evaluation, a Communication on a new strategy has not yet been produced though some development in practice is going on. The present approach is mainly consistent with recommendations of ETAN Group report *"Options and Limits for Assessing the Socio-Economic Impact of European RTD Programmes"*[3].

This set out a schematic approach:

[3] Airaghi, A. /Busch, N. /Georghiou, L. /Kuhlmann, S. /Ledoux, J.M /van Raan, A. /Viana Baptista, J. (1999): Options and Limits for Assessing the Socio-Economic Impact of European RTD Programmes. Report to the European Commission

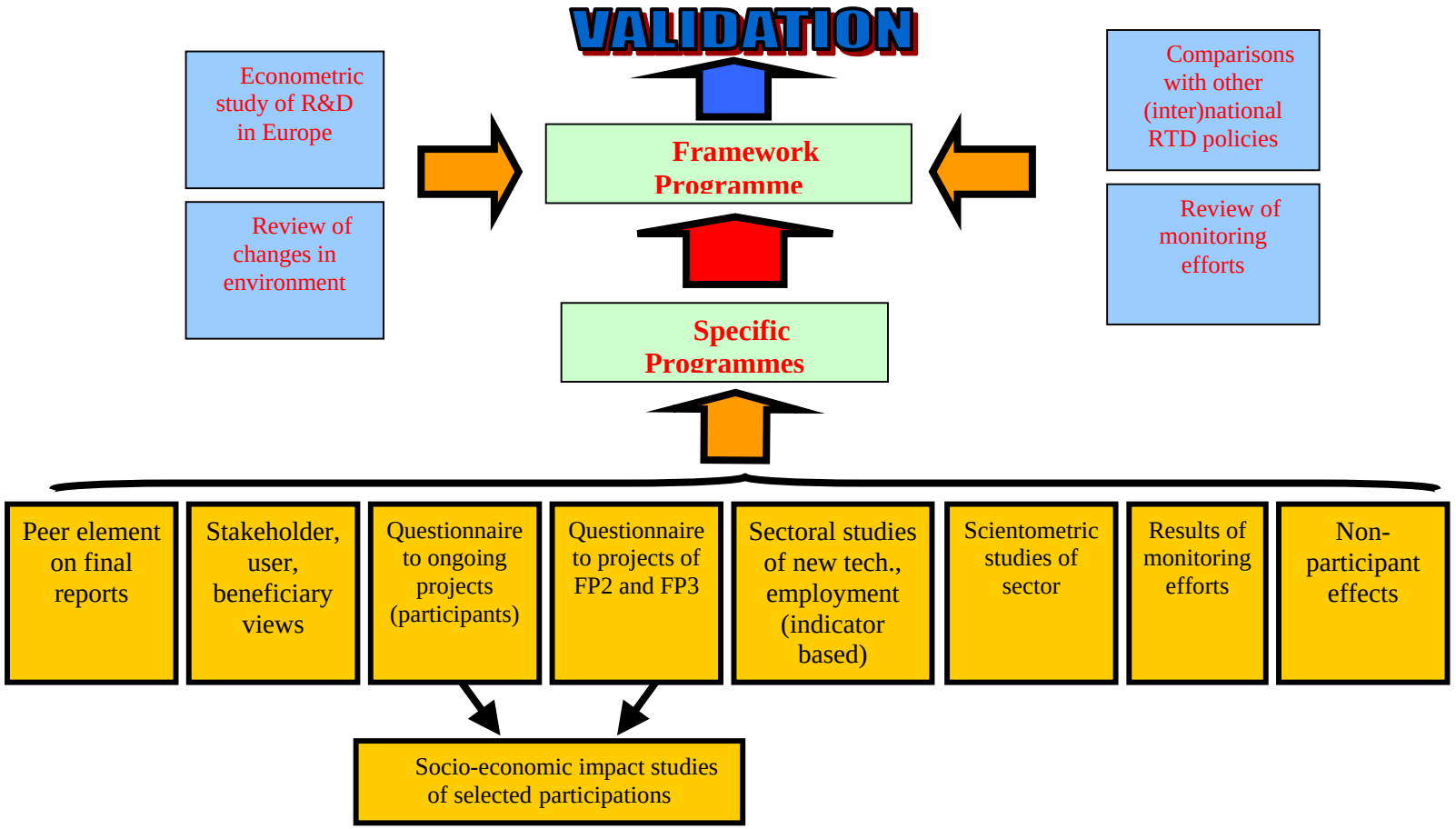

Figure 1 Schematic Approach to the Evaluation of the Framework Programme (Source: Airaghi et al)

The basic model for five-year assessments has specific programmes/key actions assessed by a panel supported by:

- Peer element in review of final reports and analysis of monitoring reports
- Stakeholder, user, and beneficiary views collected either through direct contact by the panel or more extensively through a supporting study. The aim is to identify the relevance of the programme for this group. Examples could be industrial, policy/regulatory, or social groups depending upon programme objectives
- Questionnaire sent to ongoing projects (participants) focus on process issues, strategy and management as effects not clear at this stage)
- Questionnaire to participants in previous two Framework Programmes – short instrument focused entirely on economic and social effects
- Above used as filter to select high impact project for socio-economic impact studies of selected projects/participations using interviews or elaborated questionnaires (focus on outputs for most recent completed programme and focus on outcomes/impacts for earlier programme)
- Sectoral studies of new technology and employment (using indicators) to establish landscape in which programme is operating as aid to judge appropriateness of research strategy for that sector
- Scientometric studies of sector using citation performance, co-authorship, citations to patents and mapping of research landscapes (most relevant to scientifically-oriented programmes)
- Optionally assessment of positive and negative externalities arising from programme eg diffusion of technology or displacement of competitors. Can also analyse subsequent behaviour of highly rated failed applicants.

For the overall evaluation of the Framework Programme further approaches can be used to add to the synthesis of evaluations and monitoring efforts by using

- Review of changes in landscape affecting RTD in Europe, including policy, economy, industrial, social, legal and regulatory
- Econometric study of R&D in Europe for meso and macro level impact
- Comparison with other national and international RTD policies to benchmark (with caution) administrative efficiency and priorities

This system has continued to form the basis for assessment of the Framework Programme though there has been one important change which introduces a new system for annual monitoring with one panel replacing all previous ones and linked to follow up of the Annual Management Plan.

Nonetheless, problems exist with the current approach. It is considered to rely excessively upon questionnaires to participants and the questionnaires continue to be given more prominence in reporting than the simple filter role envisaged by the ETAN panel. The problem is exacerbated by a continuing decline in response rates which call into question the validity of the resulting data.

Thinking is continuing on how to develop new approaches, particularly in the light of the broader range of instruments now being used in the Framework Programme. Two

studies by overlapping teams of European experts have focused in particular on new or more theoretically grounded approaches to evaluation:

Assessing the Socio-economic Impacts of the Framework Programme, Georghiou L, Cameron H and Rigby J (eds) 2002 http://www.les.man.ac/prest and

RTD Evaluation Toolbox - Assessing the Socio-Economic Impact of RTD-Policies Gustavo Fahrenkrog, Wolfgang Polt, Jaime Rojo, Alexander Tübke, Klaus Zinöcker (eds) 2002 http://epub.jrc.es/docs/EUR-20382-EN.pdf

The second of these studies is reviewed in detail in section 2 of this chapter.

2.3 Use of Evaluation

The only specific study of the use of evaluations of European Union RTD is now quite old though its main conclusions are still used and have been reflected in the EPUB section on this topic (see Section 2)[4].

However, a recent study by Technopolis France, *Use of Evaluation in the Commission Services* (October 2002)
http://europa.eu.int/comm/budget/evaluation/pdf/use_of%20_evaluation_final_report.pdf
has looked at the issue across all areas including some reference to RTD. Key findings included:

- Direct take up and use of results of evaluation occurs but is not the norm
- Evaluation use is more likely to be cumulative across evaluations
- Evaluation is more likely to be influential than sole cause of subsequent action
- The process of evaluation is seen as a useful opportunity for reflection and clarification of frameworks
- There is no single model of good practice
- The nature and degree of use of evaluation appears more determined by overall organizational arguments for dissemination, cross-DG consultation and routine liaison between those in the evaluation function and operational policy colleagues

The report also presents a case study of use of evaluation in DG INFSO the Directorate-General responsible for the Information Society Technologies component of the Framework Programme. From the research side of the DG's activities an examination is made of the use of the monitoring exercise of the IST Specific Programme. The Technopolis report is critical of the fact that:

> *"all too often the experts come up with conclusions and recommendations that are too general to be followed up, or that are beyond the mandate or reach of senior management of the IST programme, which, according to the terms of reference of the monitoring exercise, are the principal audience of the report."*

[4] Georghiou L, Cunningham P and Barker K, Impact and Utility of European Commission Research Programme Evaluation Reports, EC, 1990 (published by EC Monitor Programme)

To support this view they analyse the recommendations made and note that many are at a strategic level that goes beyond the scope of the Framework Programme and certainly beyond the Terms of Reference of the monitoring exercise. As a result these recommendations are not used. On the other hand, recommendations relating to operational management issues are viewed as useful and mostly implemented.

2.4 Management of Evaluation

DG Research has a long-established Evaluation Unit which has acted over the years both as a focus for methodological and strategic development of evaluation, with initiatives such as the SPEAR research programme and as the node for a succession of networks of evaluators and responsible officials), and as the unit responsible for the practical implementation of the evaluation of the Framework Programme as a whole and for those sub-programmes within the ambit of DG Research. The Unit is currently called Planning, Programme Evaluation and is part of Directorate A – Coordination of Community Activities. There is a new Unit called Impact Assessment which is concerned with ex ante evaluation.

DG INFSO has moved from a dispersed approach to a central evaluation unit since 1999. This is a small unit with six professional staff. According to the Technopolis report, people work in pairs with a 'lead' and a 'back-up' for each dossier. It states:

"The idea that someone takes an evaluation through from conception to the end. In order to improve the quality of information available to evaluators, the unit also conducts studies and provides input and advice on the identification of indicators and data collection activities."

The most important interface for the Unit is with programme managers and evaluations are timed where possible to feed into key decision points. About 90% of resources are deployed on evaluations that meet legal obligations but the proportion of self-initiated studies is increasing.

External scrutiny of the evaluation approach for the Framework Programme is carried out by an Evaluation Sub-Committee of CREST (Committee for Scientific and Technical Research) the Commission and Council's advisory committee (as agreed at the meeting of the Research Council on 10 March 1995). The Sub-Committee consists of representatives of the Member States and the Commission.

Advice is provided on:

- the terms of reference and the procedures for selecting the monitoring and evaluation Panels, while recognising the role to be played by the Programme Committees (under Article 7 of each Specific Programme);
- the reports emanating from the Panels and the actions to be taken in the light of these reports;
- the procedures adopted for monitoring and evaluating the Specific Programmes and the Framework Programme, following a review in 1997.

- Role of Steering Committee

2.5 Education, Training and Knowledge Transfer

The role of DG Budget and DG Audit in reviewing quality and establishing a Network of Evaluators to spread best practice within the Commission has already been noted. DG Research operates an Evaluation Network from national administrations in Member States. This can be seen as a part of the European Research Area (ERA) initiative, as the development of common practices and standards among evaluators is essential if countries are to begin to integrate their research support activities.

There is little activity in terms of formal training in research evaluation. Twente University has operated a four day course for many years. This originated as an initiative in the SPEAR programme but now operates on a self-financing basis. The other long-running course is provided by PREST at the University of Manchester. This is directed primarily at those doing postgraduate courses in science and technology policy and management at the university but each year a few places are opened to administrators and others from several countries. Both courses give an introductory overview of the field but they have different emphases.

3 Evaluation Toolbox

3.1 Overview of the Toolbox

The EPUB Toolbox was compiled by a network of evaluation specialists in the European Union with financial support from the STRATA programme within the EU Framework Programme. It had been observed that EU policy-making processes, and also those in many Member States, applied evaluation in a way that very much resembled what the EPUB authors understood as "monitoring". For other steps of the evaluation process, peer reviews and expert groups have mainly been used as described above. However, the EPUB Network argued that analysis of available evaluation techniques and experiences together with methodological advances showed that the evaluation process can offer more to decision-makers than currently is made use of. To address this problem, the Network analysed the methodologies for evaluating the socio-economic impact of Research and Technological Development (RTD) policies over a two-year period.

The resulting toolbox provides policy-makers, scientists and practitioners with an overview of the main evaluation concepts and methodologies, outlines their strengths and limitations, and sets them in relation to the policy context. Emphasis is set on a practice-oriented presentation.

The Toolbox, which is web-based (http://epub.jrc.es/docs/EUR-20382-EN.pdf) is structured as follows:

- Chapter 1 presents **evaluation from a user perspective** and highlights the sometimes conflicting expectations of the different actors.
- Chapter 2 describes aspects of evaluation in four broad policy areas, i.e. **financing Research and Development (R&D), the provision of R&D infrastructures, technology transfer and the legal framework**.
- Chapter 3 reviews **eleven main evaluation methodologies**, providing their descriptions, requirements for their application and good practice examples.
- Chapter 4 presents **evaluation within a distributed network**.
- Chapter 5 explores the role of **evaluation for the policy instruments in the European Research Area (ERA)** and outlines the synergies between the different evaluation methods.
- Chapter 6 provides the **results of an expert conference**, which discussed the results of the work on this toolbox with respect to the future policy context.
- The concluding remarks in the last chapter give some indications on **how to match evaluation methodologies and policy-instruments.**

Table 1 indicates the range of methodologies contained in the Toolbox. Eleven main categories are indicated with strengths and weaknesses summarised and good prac6tice examples listed. Four of these (Micro and Macro methods, Productivity studies and Cost Benefit Analysis(CBA)) are rooted in economic or econometric approaches. The first three of these generally involve statistical estimation over relatively large data sets. CBA is mainly an ex ante tool and usually involves a mixture of quantitative and qualitative data as some factors are hard to quantify or to obtain reliable data for. Control Group Approaches and the use of Expert Panels and Peer Review are tools that are specific to evaluation. In the first case the aim is to

capture the impact of policy intervention and in the second to address scientific merit (though terms of reference may go beyond this). Case Studies and Network Analysis are more generic social science research approaches but both have an important role to play in evaluation. The first is often the only way in which causality and complex situations may be addressed.

The remaining three approaches are taken from the broader set of innovation policy research approaches. Benchmarking is increasingly being used as an instrument of European research policy as part of the ERA initiative but generally has been conducted at the level of the state, though with a focus on individual policy measures in some instances – notably in the case of industry-science linkages. Foresight is a popular tool in Europe both at national and regional level. Along with Technology Assessment its use in an evaluation context is generally to provide a contextual view on issues such as whether an appropriate set of technological choices have been made for a programme of research. Finally in this category is the Innovation Survey. This is again carried out at national level, targeted at a sample of firms according to an agreed comparative framework and provides contextual information for evaluations (eg the general level of university-industry cooperation) and some indication of effects at a broader level.

In Table 2 the same methodologies are considered in terms of their level of analysis, areas of application, and which outputs, outcomes and impacts they address. An output in this case is what is produced by the R&D projects, an outcome an area of effect resulting from the outputs, either in the research-performing organization or beyond, and an impact reflects the broader area, in particular whether the method addresses competitiveness or other policy goals.

3.2 Evaluation Methodologies

Table 1 Evaluation Methodologies 1

Methodology	Ex-ante Monitoring/ Ex-post	Data Requirements	Qualitative/ quantitative	Strengths	Limitations	Good Practices
Innovation Surveys	Monitorin Expenditures g Ex-post	Micro data Profits Patents, Innovation	Semi-quantitative Quantitative	Detect trends in Innovation, soft side of innovation.	High Cost Data Demanding	Analysis of innovation process using data on the EU Community Innovation Survey
Micro Methods	Monitoring Ex-post	Micro data Expenditures Profits Patents	Quantitative qualitative categorical data	Results based on explicit formulation of theory based causal relationships R&D Additionality Control for different effects: firm size, expenditures, innovation capacity	Quality of data Persuade participant and non participant entities to disclose information Only private rate of return to R&D	Effects of public R&D subsidies on firms in West Germany Evaluation of the ITF Programme FlexCIM Effects of R&D subsidies in Spain Promotion of AMT technologies based on Swiss Micro data
Macro Methods	Ex-ante (simulation) Monitoring Ex-post	R&D Expenditures R&D output Macroeconomic data	Quantitative modelling methodology	Social Rate of return to R&D R&D Spillovers Long term impact	Average returns Robustness of results Time lags for observation of the effects	Modelling approaches: OECD Interlink, IMF Multimod, EU Quest. R&D Spillover studies: Jaffe, Nadiri International spillovers: Eaton and Kortum Mohnen, Evenson
Productivity Studies	Monitoring Ex-post	Micro data Expenditures Profits R&D, Patents	Quantitative modelling methodology	Estimation of effect of R&D on productivity Rates of return to R&D	Quality of data Measurement of stocks	Productivity studies (Van Ark) Growth accounting (Griliches, Jorgenson) Micro datasets (French INSEE) and US Census of Manufacturers
Control group approaches	Ex-post	Micro data Expenditures Profits Patents	Quantitative	Capture the impact of policy intervention on the participant entitity	High Implementation Cost Data Demanding	Collaborative industrial Research between Japan and US Evaluation of RTDI instruments in Ireland Participation of Ireland in Eurpoean Space Agency
Cost Benefit Analysis	Ex-ante (specially) Monitoring Ex-post	Micro data Profits Costs	Quantitative	Socio-economic effect of intervention	Judgement technique Subjectivity Not generalisable	US Advanced Technology Programme US National Institute of Standards Methodology
Expert Panels /Peer Review	Ex-ante Monitoring Ex-post	Project programme data	Qualitative Semi-quantitative	Evaluation of scientific merits Flexibility Wide scope of application Fairness	Peers independence Economic benefits not captured	Evaluation of Large Infrastructures Evaluation of EU Programmes
Field /Case studies	Monitoring Ex-post	Project programme data	Qualitative Semi-quantitative	Observation of the socio-economic impacts of intervention under naturalistic conditions Good as exploratory and descriptive means of investigation Good for understanding how contexts affect and shape impacts	Results not generalisable	Telematic innovation in the health care sector. Evaluation case studies reviewed in Georghiou and Roessner (2000)
Network Analysis	Ex-post	Project programme data	Qualitative Semi-quantitative	Comprehensive empirical material compilation for policy purposes Cooperation linkages	Time involved in collecting the survey information Persuasion requirements	RTO systems Interdisciplinary centers of medical research
Foresight/ Technology Assessment	Ex-ante Monitoring	Qualitative data Scenario	Qualitative Semi-quantitative	Consensus building to reduce uncertainty under different scenarios Combination on public domain and private domain data Articulation and road mapping of development of new technologies	Impossibility to detect major RTD breakthroughs	Benchmarking of I ISI/FhG capacities against Foresight results

Benchmarking /Ranking	Ex-ante Monitoring	Science and technology Indicators	Semi-quantitative	Comparison method across different sectors	Data detail requirements Not trasnferable	EU Benchmarking national policies Innovation Trend Chart

Table 2 Evaluation Methodologies 2

Methodology	Data application level	Areas of application	Output	Outcome	Impact
Innovation Surveys	Firm Industry Economy-wide	Innovation	New products and processes Increase in sales Increase in value added Patent counts, IPRs	Creation of new jobs Innnovation capacity building	Enhanced Competitiveness Institutional and organisational efficiency, Faster diffusion of Innovation Employment
Micro Methods	Plant Firm Industry Economy-wide	Sectoral Returns to R&D	Output and value added (collect baseline info for before-after comparisons)	Sectoral productivity industry sectoral spillovers	Firms competitiveness
Macro Methods	Firm Industry Economy-wide	Sectoral Regional Economy-wide	Output and value added	Change in R&D Capital, Human capital, Social capital International R&D Spillovers	Regional, country productivity Employment, Good governance Economic and social cohesion
Productivity Studies	Firm Industry Regional Economy-wide	Sectoral Regional Economy-wide	Output and value added	knowledge, geographical and International R&D Spillovers	Regional, country productivity Employment Economic and social cohesion
Control Group Approaches	Firm Industry	Technology implementation Innovation	Output and value added (on supported and non supported firma)	Additionality Rate of return to R&D	Firm, industrial competitiveness
Cost Benefit Analysis	Firm Industry	Health Environment Energy Transport	Value added benefit-cost ratio IRR Consumer surplus	Health improvements Consumer protection Environmental sustainability	Quality of life Standard of living
Expert Panels/ Peer Review	Firm Industry Economy-wide	Scientific merit Technological capacity	Publication counts Technological output	Scientific and Technological capabilities	R&D performance
Field/ Case Studies	Firm Industry	Science-industry relationships	Detailed inputs and outputs	firms RTD capabilities on the job-training educational schemes	Industrial competitiveness Quality of life Organisational efficiency
Network Analysis	Firm Industry Regional	RJVs, cooperation science industry Clusters	Cooperation linkages	Cooperation in clusters Social embededness	Efficiency of institutional relationships
Foresight/ Technology Assessment	Institution Regional Economy-wide	Technology Trends	Identification of generic technologies Date of implementation	Technological capacities	Technological paradigms shifts
Benchmarking /Ranking	Firm Industry Economy-wide	Efficiency of technology policy	S&T indicators	Technology capabilities	Industry competitiveness Good governance

3.3 Matching Method to Policy Measure

Table 3 summarises the suitability of the methods for different policy instruments. It may be observed that the area of technology transfer and diffusion is addressed by the broadest range of methodologies, while the rather specialised European instrument, the Network of Excellence has the smallest range of suitable approaches recommended.

Table 3 Evaluation Matrix – matching policy needs and methods

Methodology	Data application level	Areas of application	Output	Outcome	Impact
Innovation Surveys	Firm Industry Economy-wide	Innovation	New products and processes Increase in sales Increase in value added Patent counts, IPRs	Creation of new jobs Innnovation capacity building	Enhanced Competitiveness Institutional and organisational efficiency, Faster diffusion of Innovation Employment
Micro Methods	Plant Firm Industry Economy-wide	Sectoral Returns to R&D	Output and value added (collect baseline info for before-after comparisons)	Sectoral productivity industry sectoral spillovers	Firms competitiveness
Macro Methods	Firm Industry Economy-wide	Sectoral Regional Economy-wide	Output and value added	Change in R&D Capital, Human capital, Social capital International R&D Spillovers	Regional, country productivity Employment, Good governance Economic and social cohesion
Productivity Studies	Firm Industry Regional Economy-wide	Sectoral Regional Economy-wide	Output and value added	knowledge, geographical and International R&D Spillovers	Regional, country productivity Employment Economic and social cohesion
Control Group Approaches	Firm Industry	Technology implementation Innovation	Output and value added (on supported and non supported firma)	Additionality Rate of return to R&D	Firm, industrial competitiveness
Cost Benefit Analysis	Firm Industry	Health Environment Energy Transport	Value added benefit-cost ratio IRR Consumer surplus	Health improvements Consumer protection Environmental sustainability	Quality of life Standard of living
Expert Panels/ Peer Review	Firm Industry Economy-wide	Scientific merit Technological capacity	Publication counts Technological output	Scientific and Technological capabilities	R&D performance
Field/ Case Studies	Firm Industry	Science-industry relationships	Detailed inputs and outputs	firms RTD capabilities on the job-training educational schemes	Industrial competitiveness Quality of life Organisational efficiency
Network Analysis	Firm Industry Regional	RJVs, cooperation science industry Clusters	Cooperation linkages	Cooperation in clusters Social embededness	Efficiency of institutional relationships
Foresight/ Technology Assessment	Institution Regional Economy-wide	Technology Trends	Identification of generic technologies Date of implementation	Technological capacities	Technological paradigms shifts
Benchmarking /Ranking	Firm Industry Economy-wide	Efficiency of technology policy	S&T indicators	Technology capabilities	Industry competitiveness Good governance

3.4 Good Practice Steps

The Toolbox offers a detailed range of pointers to help in the tasks associated with the practical steps involved in designing, implementing and using evaluations. These are listed below:

Evaluation planning

- Provide an early and adequate scheme for the evaluation design and integrate it into the policy intervention design to ensure that intervention objectives are clearly defined and can be effectively evaluated.

- Base the public intervention on a demonstrated market or systemic failure, which the intervention should solve.
- Define requirements on data compilation and updating during the intervention design stage. Ex-post evaluation will critically depend on the quality of compiled data.
- Introduce new methods in ex-ante evaluation that favour diversity and the taking up of new risks and multidisciplinarity. Peer review is a significantly conservative approach in the evaluation of research proposals, risky projects are likely to get worse scores from peer review. Mainstream science is better positioned when adopting peer review methods.

Operational and management issues

- Allocate sufficient time and monetary resources to evaluation. This is justified as the aim is to ensure that public money is efficiently and wisely spent.
- Promote independence to ensure credibility of results, for this purpose it might be relevant to use external evaluation experts (from other countries).
- Involve policy makers and project managers in the evaluation so that their perceptions and priorities are fed into the evaluation design and during the evaluation execution.
- Separate in evaluation the strategy function from the operational function. Evaluation as a demonstration of impact is only one input to strategy definition.
- Strengthen transparency by publishing the terms of reference, criteria's used in the evaluation and disseminating the produced evaluation results to a broad audience of interested bodies.

Evaluation priors

- Clarify the implicit policy rationale of the intervention when conducting an evaluation. Identify the objectives of the policy intervention being evaluated, discussing the intervention logical framework, including implicit assumptions and establishing the feasibility of evaluating them.
- Define the intervention jointly with concrete targets that will facilitate the evaluation of the instrument, e.g. "increase the publications in the field of genetic technology by 20 per cent or increase productivity by 10 per cent".
- Ensure the compilation of data before and after the intervention as well as on supported and non supported units to allow to control for the counterfactual.

Method implementation

- Adapt methodologies to deal with the particular evaluation requirements and to answer relevant questions. Evaluation should not be perceived as mechanical process.
- Definition of objectives determines the methodology selection.
- Combine different methodologies and different levels of data aggregation to improve the understanding of the multidimensional effects of the policy intervention.
- Incorporate systemic considerations into evaluation as science and technology is likely to modify institutions structure and behaviour.
- Separate when possible the evaluation of the scientific merit provided by traditional established methodologies such as peer review from the evaluation

of the other socioeconomic objectives provided using the support of expert panels.
- Evaluate the profile of supported and non-supported firms including those rejected and those who did not apply for support. Control group approaches are especially valuable in this context.
- Establish intended and unintended effects of the intervention. Analyse failure as well as success histories.

Strategic evaluation
- Integrate RTD evaluation practices with other sources of "distributed intelligence" such as technology assessment and foresight to support strategy policy definition.
- Develop criteria to evaluate the increasing strategic role of systems and institutions in science and technology. New applications of benchmarking, foresight and network approaches could be used to evaluate increasingly relevant topics such as institutional capacity, behavioural additionality and networking.

Dissemination of evaluation results and recommendations
- Broaden the use of evaluation results by incorporating the views of the potentially interested audience such as industry, target groups and social communities representatives.
- Introduce the requirement that programme managers report on implementation of recommendations made in the evaluations.
- Produce timely evaluation reports and in a clearly and understandable language to increase impact.

3.5 Meeting User Needs

When considering the use of evaluation the Toolbox states that it is common to differentiate between 'process use' and use of results. Process use is increasingly recognised as an important form of evaluation use where the process of undertaking the evaluation helps parties to clarify their thinking and share information. At the final stage in the process of evaluation, there are opportunities to implement results. Successful implementation can be seen as related to the following three main factors[5]:

- **Absorbability** -The dissemination of the content and findings of evaluation studies requires good levels of awareness of the study among the audience, particularly where it combines various elements. The final report itself must be digestible and target its recommendations at an appropriate level: not too specific and not too general. It should be delivered in time for follow-on decisions, it may need later validation, and may have objectives linked to its timing and relevance to programme and policy cycles.
- **Credibility** - The credibility of evaluators should be taken into account in awarding the original tender, and should not relate to technical ability and proven quality and cost effectiveness, but to fairness and independence, and reputation. The results produced should be of high quality and founded on solid evidence and sufficient depth and breadth of coverage.

[5] Georghiou L, Cunningham P and Barker K, Impact and Utility of European Commission Research Programme Evaluation Reports, EC, 1990 (published by EC Monitor Programme)

- **Steerability** - Some policy and programme initiatives are more open to being 'steered' by evaluation than others. Evaluation is only one among many influences on policy decisions. In some policy domains a high proportion of implements is located with policy makers and politicians and in others less so. The extent to which evaluations can make a contribution towards programme development and policy decisions is a factor that needs to be taken into account when decisions are made about funding evaluations. What is the added value of an evaluation also needs to be considered in these terms.

A more recent consideration of meeting the needs of evaluation users is summarised in the Toolbox as Table 4. The table points out the gaps in expectations between evaluators and users of evaluations.

Table 4 The Customer Gap

What evaluators want	What policymakers say
Clearly defined and hierarchical objectives	Programmes are a compromise involving multiple and conflicting objectives
Guaranteed independence	Recommendations must be within realistic policy constraints
Time and resources to do the job	We need the results in three months
Full access to information and stakeholders	Everyone is overworked and too busy

Source: Georghiou 2002[6]

[6] Evaluation in the Balance: Matching Methods and Policy Needs, Conference Belgian Presidency, Brussels 2001 International best practices in evaluation of research in public institutes and universities

4 Application of the EU Framework Programme Approach and the Concept of the Toolkit to Japan

4.1 Introduction

In this part of the paper, the concepts, methods and practices in the European Union described in Sections 1 and 2 will be discussed in relation to their applicability to the Japanese situation. The main headings used will be methods, management of evaluation, use of evaluation and education of evaluators. Before beginning the discussion a brief summary is given of the existing status and practice of research evaluation in Japan.

4.2 Overview of Japanese Evaluation Approach and Issues

Evaluation of research and development in Japan needs to be seen in the context of the 1997 First National Guidelines and the 2001 Revised National Guidelines. The framework established in the first document was extended in the second with an emphasis on three main groups of factors:

- the need to upgrade fairness and transparency (using objective indicators and external evaluation, and promoting rapid diffusion of results);
- strengthening the link with budgeting and human resource allocation; and
- establishing a resource base for evaluation (including training for evaluators).

Other issues raised include the key questions of what to evaluate (measures, themes, organizations, researchers), and when to evaluate (ex ante, mid-term, ex post, tracking programmes).

General criteria for evaluation are set out:

- Necessity – Scientific and technological significance, socio-economic significance, appropriateness of objectives
- Efficiency – Appropriateness of implementation plan, organization, output vs input
- Effectiveness – targeting and achievements, creation of knowledge or socio-economic benefits, human resource developments

Another key issue raised is minimising the burden of evaluation upon managers and researchers.

Specific guidelines for the Ministry of Economy Trade and Industry and its agents have also been developed with the most recent version being the 2002 revision for METI and a 2001 revision for NEDO (New Energy and Industrial Technology Development).

METI itself has broad scope for its evaluation activities covering the evaluation of organizations, field evaluations (for example energy or environment), evaluations of the R&D Grant system and of small programmes, plus what is described as project package evaluation (analogous to programme evaluation with multiple projects and to include regulation and standards issues), and finally project evaluation[7]. NEDO has a

[7] Naomichi Miyazawa, R&D Evaluation in METI and NEDO, http://www.meti.go.jp/policy/tech_evaluation/english/pdf/f00/2002/f0001630.pdf

much more restricted scope, being involved directly only at the level of project evaluation.

Project evaluations in the METI context are generally carried out as interim evaluations – in the middle year, and ex post – in last year or following year. In some cases there is then a follow-up evaluation several years after. The basic approach is that they are carried out by external panel review with the relevant promotion division and the evaluation division providing the secretariat. In the case of follow up evaluations METI is examining 2-3 projects per year to investigate industrial and social effects. The method is based upon four steps:

1. Interviews with relevant persons covering:
    - background of participation in the project
    - Impact on science and technology field
    - Effect on research and development faculty in the laboratory/company
    - Economic impact on the industry
    - Effect on standard of life and society
    - Feedback to policy planning
    - Suggestions to the project management

2. Collection of project-related materials

3. A panel review by external evaluators. The composition of the panel is:
    - Experts in the scientific/engineering fields of the project
    - Experts in R&D management
    - Economists
    - Users of results of the project
    - Commentators from mass media

4. Production of evaluation report with attributed comments

Issues raised as a result of these evaluation experiences include many of those addressed in this paper such as training of evaluators, establishing a database of results, simplification of the process to reduce the burden, and how to improve socio-economic impact evaluation.

4.3 Appropriate Methods for Japanese Programmes

The basic methodological challenge for the European Commission and METI is the same – that of how to improve socio-economic impact evaluation. Despite the much longer evolution of the Commission's evaluation system, and numerous analyses calling for more information of this type, the basic approach in Brussels remains one of Panel review.

Before embarking upon a comparison it is worth remembering that in many ways as a sponsor and customer of research the Commission corresponds to a whole government not a single Ministry. Thus a substantial proportion of its research is intended to support policy and regulation in areas such as health, safety and the environment. It also has a mission to improve European science more generally

(through schemes such as the Marie-Curie mobility actions). While METI does support research with broader policy goals such as energy supply and conservation, the closest comparison is with those parts of the Framework Programme aimed at support for industrial competitiveness. The goals cannot be completely separated since research in support of a policy objective can also lead to industrial competitiveness in related products (eg health technology) but the order of priority is clear.

A comparison of actual current practice in methods shows a basic similarity, that is the use of **panels of independent experts** relying upon interviews, surveys and inspection of project materials. The **criteria** applied are also very similar though phrased differently.

In terms of **scope of evaluation** there is a difference in the units addressed. For the Commission the sub-programme, or priority area, has formed a natural unit for evaluation (with a legal requirement attached). Budgets are allocated in this way and work-programmes written against which objectives-achievement may be assessed. The definition of a programme is not the same for METI (in European terms many METI sponsored projects would be seen as programmes. For METI the term programme is used to encompass a wider aggregation including and broader scope of issues to be addressed. However, the link with a budgetary focus and objectives is less clear as a result.

Both the Commission and METI carry out some evaluations outside the accountability driven frame of projects and programmes. For the Commission these included the national impact studies and more recently an evaluation targeted at the new policy instruments. For METI the main variant is that of field reviews (more common at national level in Europe).

Moving to the core issue of socio-economic evaluation, the first area to be examined is that of the **ETAN report** since this continues to form a model for practice. Despite efforts to broaden the approach (partly inspired by falling response rates) the questionnaire remains the most used instrument. Increasingly the Commission has been rec9ognising the value of detailed case studies, both for evaluation purposes and to use as "success stories" in helping to gain support for programmes. The two applications clearly need to be kept separate, with success cases only being prepared after evaluation.

In general, the broader approaches recommended by the ETAN panel have been carried out by the indicators unit of DG Research but for other purposes. Evaluations occasionally make cross-reference to them but have only commissioned work of this kind on an experimental basis. The same may be said of failed applicants. A particular problem exists here in building databases because of data protection issues.

In a sense, the EPUB Toolbox (along with the ASIF study) represent the Commission's efforts to move in the direction of more rigorous and methodologically grounded approaches. The Toolbox methods will be considered in relation to METI practice in the order in which they occur in the tables.

Innovation surveys

The information gathered via an innovation survey (in the same basic format as the Third Community Innovation Survey) is available in Japan (ie Japan National Innovation Survey 2003). At this level mainly contextual information for programme and project evaluations can be gathered. However, this is useful in establishing facts such as the proportion of firms by sector that participate in government programmes or the extent to which firms access knowledge from universities or government laboratories. These behaviours can be correlated with innovative performance. It can also establish a benchmark for innovative activity that may be used to assess the overall effect of government efforts to stimulate innovation or to compare firms attitudes to different forms of intervention – for example fiscal incentives versus grants.

Micro methods

Microeconomic methods generally involve a statistical comparison on a before/after, counterfactual or else with a similar group of firms not experiencing the policy instrument in question. These methods have not been applied to EU R&D programmes to date but they have been used in national situations in Europe. They are at their strongest when applied to diffusion and extension programmes where there are large numbers of similar firms experiencing a similar stimulus. Hence the successful application to the Swiss AMT programme. However, more recently they have also been applied to R&D programmes in Germany and also in a specific study in Japan (cited in the EPUB Toolbox) Sakakibara, 1997[8], who examined the relationship between participation in MITI-sponsored R&D consortia and levels of R&D expenditure, research productivity and knowledge spillovers. It may be concluded that the state of knowledge in Japan in this area of evaluation is at an equivalent level to Europe.

Macro methods

Evaluation at the macro-economic level is intended to measure the global socio-economic impact in the long term, that is to say the benefits to society as a whole rather than the benefits to the programme participants. The strength of the approach is this broad scope measurement while its weakness is that the data it treats are affected by a wide range of factors. The approaches are best used in an ex ante context to make the broad case for R&D support policies. They have no function in the evaluation of a specific programme. A good example is the work of Guellec and van Pottellsberghe[9] who estimate the contribution of various sources of knowledge (R&D capital stocks performed by the business sector, by foreign firms, and by public institutions) to productivity growth as well as the determinants of privately funded and performed R&D. A particularly useful outcome is the light such approaches can shed on policy-mix issues. For example, the estimates show that government funding of business R&D is substitute to fiscal incentives, complementary to university research, and does not interact with government research. In other words, increasing the direct funding (tax incentives) of business research reduces the stimulating effect of tax incentives

[8] Sakakibara, M. (1997): Evaluation of Government-Sponsored R&D Consortia in Japan, in: OECD (ed.), *Policy Evaluation in Innovation and Technology – Towards Best Practices*, Paris.

[9] Guellec D. and van Pottelsberghe B. (1999), Does government support stimulates private R&D ?, *OECD Economic Studies*, 29, 1997/II, 298-122.

(direct government funding). In addition, increased government funding of business research appears to reduce the negative effect of university research on business funding, possibly because government funding helps firms to absorb knowledge from universities that may be poorly used.

With respect to application in Japan the Toolbox cautions certain conditions for successful use of macro-methods:

- the availability of a large scale socio-economic dataset;
- a high degree of expertise;

enough time to build the model prior to any evaluation exercise;

- the implementation of policies which do not lead to a high dilution of economic effects.

Of course, there is substantial Japanese experience in macro-modelling of the economy including some studies by ESRI, NSTEP and other which include R&D as a key variable. The issue is whether these have been used in c complementary manner to help interpretation of evaluation findings using meso and micro level methods.

Productivity studies

The EPUB report states that productivity refers to the amount of output that a given set of inputs can produce. The ultimate purpose of RTD activity is usually to raise productivity. Hence productivity measurement is central to the *ex post* evaluation of the success or failure of such policies. The *ex ante* evaluation of a proposed RTD activity in most cases would hinge ultimately on its effect on productivity. Connecting the use of a particular instrument or intervention with the ultimate effect on macro or meso level productivity may often prove extremely difficult. Too many other factors affecting productivity vary to allow for appropriate controls. Hence the effect of many interventions may be better observed at the micro level, especially if an experiment is undertaken involving a control group of production units. At the micro level any intervention that affects some producing units more than others can in principle be evaluated in terms of its effect on the measured productivity of those units. Approaches used generally involve estimating a production function using the assumptions of growth accounting or production function econometrics. Once again, studies of this kind have already addressed the Japanese situation. For example Branstetter and Nakamura (2003)[10] have investigated changes in the output and productivity of research and development activities in Japanese manufacturing firms over the 1980s and 1990S and drawn a number of general research and innovation policy conclusions.

Control group approaches

The use of control groups in programme and instrument evaluation is a means of addressing the problem of how to exclude the effects of variables external to the sphere of influence of the instrument being evaluated.. To be rigorous, it requires statistically significant sample sizes, stratified exactly the same as the test population. It may be that such samples do not exist for a particular application, either in terms of numbers or exact comparability of objectives. The use of control groups in this

[10] Branstetter L and Na,kamura Y, Is Japan's Innovative Capacity in Decline? National Bureau of Economic Research Working Paper 9438

context must not be regarded as having the accuracy or precision of laboratory experiments nor of clinical trials in the pharmaceutical and healthcare industries. Within the evaluation context it is particularly useful as a measure of the additionality of public support.

A typical approach is to establish three groups of actors:

> **Group 1** consists of beneficiaries of the instrument
>
> **Group 2** consists of those who applied unsuccessfully for funding, but who completed the project, using their own funds or other support instruments such as venture capital or a loan. Ideally, the only difference between Group 1 and Group 2 is that Group 2 did not receive funding under the instrument being evaluated, but on average Group 1 are likely to be more experienced in submitting proposals and in carrying out R&D. The original Group 2 project may have been slightly modified or executed slower, and in fact the different funding mechanism may have changed aims and objectives as well.
>
> **Group 3** consists of those who did not seek funding under the instrument, but executed a comparable project using their own funds or other means of support. They will clearly be most useful in gaining insights into the management and promotion of the instrument, and its relevance to the needs of the target clientele.

Groups 2 and 3 are control groups. The principal practical difficulty involved in this method consists first in identifying the members if the control groups and secondly in persuading them to cooperate. An attempt by PREST to use failed applicants as a control group in an evaluation of a METI programme failed because only a very small number could be identified because records are not kept of applications. That study concluded that this method was unsuitable for use in Japan in the present circumstances. Difficulties have also been encountered in Europe as a result of data protection issues. However, the additionality issue is a core aspect of evaluation of government sponsored programmes and perhaps has been under-emphasised in past Japanese evaluations.

Cost benefit analysis

Cost benefit analysis (CBA) is widely used in the assessment of investment projects. CBA when applied to public projects addresses social as well as private costs and benefits and thus can provide a full assessment of economic efficiency and the possibility of comparison between investment choices. It seeks to take into account both direct effects and indirect effects or spillovers. Results are typically portrayed as Net Present Value, Internal Rate of Return or Cost-Benefit Ratios. In such calculations the choice of discount rate is of particular importance. Despite its widespread general use in ex ante evaluation CBA is rarely used for the assessment of R&D programmes. There is no European example given in the Toolbox but a study from the US Advanced Technology Programme indicates that CBA has value in encouraging programme managers and participants to be more systematic in their consideration of costs and benefits and to become aware of a wider range of benefits. The principal limitations of the approach lie in attributing effects, especially after significant time has elapsed, and in the intangible nature of some effects. The Toolbox

concludes that the method is most appropriately applied to large scale mission oriented programmes such as energy, space and aeronautics and is unlikely to be effective in evaluations programmes where there are larger numbers of smaller projects.

Expert panels/ peer review

The terms peer review and panel review are quite different. Peer review refers to the judgement of quality of work by experts in the same scientific field. In reality it only addresses the issue of scientific quality but peers are often asked also to comment on broader issues such as the management of research programmes. Panel review takes the basic format of peer review and extends it by adding expertise in other aspects, such as economists, evaluation experts, or by adding stakeholders such as representatives of industry or other potential users of results. It is sometimes known as modified peer review and sometimes as merit review.

There are many ways of organising panel review and efforts are generally made to demonstrate the independence of members and to control the dynamics of how they interact with each other. In terms of answering questions about socio-economic impact it is generally insufficient for a panel simply to inspect final reports of projects. Good practice involves equipping the panel with information from supporting studies which they have played a part in commissioning in the first place (see comments about ETAN report above). This of course implies that at least one of the panel members is able to interpret, for example, econometric studies. Direct engagement of the panel in interviews with participants and stakeholders is also seen as desirable.

For the European Commission, the panel is necessary for another reason – it allows members from several countries to take part in the evaluation. While this aspect does not apply to Japan, the presence of a wider range of stakeholder groups does add to the legitimacy of the findings. This leads to a further comment on panels – they often consist of prestigious members and hence their views are more likely to be taken into account. An issue for Japan to consider is that of whether to place the panel above a study team or vice versa. While the European Commission always places the panel in prime position for its formal evaluations, in several member state evaluations the panel is simply one part of a broader evaluation run by a contractor and is confined to the specific issue of scientific quality. Managing this relationship has often proved problematic.

Field/case studies

The Toolbox states that field studies involve the *direct observation of naturally occurring events*. This implies that, instead of investigating behaviour under artificial conditions, such as in the laboratory or under rigorously controlled condition; and instead of telescoping the process of enquiry into a 'snapshot' of time and space, such as in an interview or questionnaire, field studies involve the prolonged and relatively uninterrupted investigation of behaviours in their indigenous social setting. As such, field studies are made up of many potential methods and techniques. A common format within which these methods and techniques are assembled is that of a 'case study'.

Field and case studies may be used in a wide variety of settings normally where an exploratory approach is called for. While criteria for selection of case studies need to be defined they rarely are conducted in sufficient numbers to form a representative sample because of resource limitations. The theoretical foundations and methodological guidelines for case study approaches come from outside the field of evaluation, being located in the general social science literature. Techniques for data collection may be unobtrusive (for example by observation) or obtrusive (for example based upon interviews). There are various analytical techniques available which can include delineation of events that may then be analysed quantitatively over time.

Within evaluation case studies provide a means by which information from a variety of sources may be assembled and interpreted. The main drawback is the difficulty of drawing generalisable conclusions. In Japan case study approaches have been used in the recent evaluation of the Basic Plan.

Network analysis

The general trend towards promotion of networking in innovation policy (for example academic-industry and industry-industry links) has led to a need for evaluations to analyse the specific characteristics and effects of networks. These approaches draw from the social sciences a special research focus – social network research –, which serves to analyse the structure of (cooperative) relationships and the consequences for actors' decisions on actions. The Toolbox argues that the premise of this approach is to be able to formulate explanations for the actions/activities of individuals by describing and analysing their social embeddedness, i.e. individual actions are neither attributed to the normative convictions of the actor nor to the mere membership of a certain category, such as e.g. age groups, but to the individually structured relationships between the actors. Networks are generally mapped using characteristics of the relationships eg strength of a bond between members, or in terms of the overall network characteristics eg the density of particular parts of the network.

Networks are generally constructed on the basis of survey data or else from pre-existing measures of connectivity such as co-authorship. In evaluative terms network analysis can be used to identify progress in the achievement of network related policy goals (or to identify corresponding weak spots). Characterisation of networks by their structure and density can allow normative conclusions to be drawn about the suitability of particular organisational arrangements. The European Commission has periodically used studies drawing upon network analysis to support its evaluations. The most recent example is a study for DG Infosoc which maps information technology research networks using a variety of inputs.

Techniques of this type are likely to be equally useful in Japan since the promotion of better interchange and knowledge flows is a central plank of policy objectives.

Foresight/technology assessment

While evaluation has traditionally been seen as a retrospective process – examining the consequences of past actions – there has been an increasing tendency in Europe to view it as complementary to the forward looking activities of foresight and technology assessment. Foresight is generally seen as going beyond technological forecasting by means of engaging a participative process in which the process itself is

considered as valuable as the output in achieving policy goals. It provides a forum in which actors can share their visions of the future and the assumptions which underpin them. Foresight can be used in a variety of ways including as the toolbox states:

- to find out new demand and new possibilities as well as new ideas,
- to identify a choice of opportunities, to set priorities and to assess potential impacts and chances,
- to discuss desirable and undesirable futures,
- to prospect the potential impacts of current research and technology policy,
- to focus selectively on economic, technological, social and ecological areas as well as to start monitoring and detailed research in these fields.

Techniques and methods to pursue foresight include Delphi surveys, construction of scenarios and brainstorming of various types.

Technology assessment aims to anticipate effects and impacts of new technologies and to feed this knowledge back into the decision-making process. It is more an advisory than an analytical approach and is increasingly reliant on participatory approaches.

The Toolbox indicates some circumstances in which foresight and technology assessment can form a useful input to evaluation:

- *Financing R&D:* foresight exercises and technology assessment can help with *priority-setting* under the condition of scarce public budgets and competition for funding – evaluation might either assess funded research and innovation activities *ex post* in the light of foresight and technology assessment results by using them as a kind of benchmark, or rank envisaged funding themes *ex ante*.
- *Provision of R&D infrastructure:* foresight exercises and technology assessment can help to evaluate the actual or envisaged *priorities of research institutes*, by using the exercises' results as a benchmark (see example of Fraunhofer evaluation, below).
- *Technology Transfer/Innovation Diffusion*: foresight exercises and technology assessment can help to identify the quality and extent of the present or future demand for research results and technological developments, i.e. for the *likelihood of successful innovation.*
- *Standards, Regulations, IPRs:* foresight exercises and technology assessment can help to *characterise the need* for technical standards, regulations, and for the appropriateness of IPR regimes, in the light of identified present or future technical, social or economic risks and potentials, thus enlightening the evaluation of related policy measures.

Japan is well advanced in the area of foresight in particular and has the potential for successful application of this mix of “strategic intelligence” approaches.

Benchmarking

Benchmarking is defined in the Toolbox as a continuous systematic process for comparing the performance of for example organisations, functions, processes of economies, policies or sectors of business against the best in the world, with the aim of improving performance and learning. Generally quantitative indicators are used. Though originated in an industrial context, the approach has spread to public sector organisations (including laboratories) and to policies. In Europe the approach received a major impetus with the launch of the European Research Area and the "open method of coordination" of RTD policies which involves systematic comparison between member states (and others). Benchmarking at policy level has proved problematic because of difficulties in establishing causal links between policies and effects. At a more disaggregated level eg scientific disciplines or infrastructures, the process has proved more useful. A particularly successful application was in benchmarking industry-science relations.

Intelligent benchmarking may be seen as a form of evaluation but one which should be approached in a cooperative manner with the principal aim being learning rather than formation of summative judgements. In this vein it should be repeated regularly. In terms of broader application in RTD policy Japan is already significantly engaged in activities similar at least to benchmarking. However, there is scope for more systematic use of international comparison in METI's programme evaluations as an enhancement to the information derived from national sources.

4.4 Management of Evaluation

In terms of evaluation management it is possible to note many similarities between the Japanese situation and that of the European Commission. First it may be noted that both operate within a legislative framework. The guidelines for the Framework Programme approach were initially in the same situation as those for Japan now, in that they were specific to RTD. However, research has now become much more integrated with the Commission's general evaluation framework. This process is still ongoing and its full consequences are not yet clear.

A second similarity is the existence of dedicated units within the service to organise and promote evaluation, along with the expectation that programme management will also support evaluations in their sphere. Reliance on panels with some external support from specialised consultants provides a third area of similarity.

There are also differences. The European Commission is accountable in multiple ways, to Member States as well as to the European Parliament. This process results in panel selection being constrained by issues of national balance as well as by expertise and stakeholder presence. This situation need not apply in Japan. There is also a tendency for the highest level evaluation panels (eg whole Framework Programme) to have a political Chairman – typically an ex- research minister. This has resulted in some reports that have tended to ignore some of their terms of reference for evaluation and instead to concentrate on future strategy. This does not mean that independence has been compromised but independence can also be used to deviate from what is required. There is no benefit to Japan in following this route. Evaluation forms a very useful input to future strategy-making but only when it is done rigorously and with a focus on its terms of reference.

The Commission also does not provide a good model in terms of resources for evaluation – this has been the subject of continuing complaint as there is no dedicated budget. In Member States it is quite acceptable for ½ to 1% of programme costs to be set aside for evaluation. This provides a better benchmark for Japan. These resources should however be deployed strategically – for example more to be spent on new or complex measures and less on routine activities where the result is predictable. Time is also a scarce resource that should be allocated as generously as possible – the quality of European evaluations has often been reduced by allowing insufficient time, especially for analysis and conclusions.

There could be benefits for Japan in considering the merits of external scrutiny of the evaluation function itself – with both results and process being examined at a higher level. Keeping evaluation transparent is also good practice. As the Toolbox suggests, publishing terms of reference and criteria are important. Agreeing the rationale and programme logic with programme managers is a useful first step to ensure that the evaluation is considered fair by all concerned.

4.5 Use of Evaluation

From what is known about the use of evaluation in the European context several lessons may be drawn which have more general significance. The first is that expectations have to be moderate – evaluation is one influence among others and is only exceptionally the sole cause of change in policy. Nevertheless the study of use of evaluation showed that evaluation has a cumulative effect – this is important as it suggests that a well ordered evaluation system consistently delivering results is more effective than isolated exercises even if the latter are high profile.

There are also relevant findings concerning the presentation and level of results. The concepts of credibility, absorbability and steerability all imply that evaluations must be realistic in the way they enter the policy system. Recommendations need to be cast within reasonable grounds of possibility for implementation and the presentation itself must be both clear and properly supported by the evidence. The more success there has been in accessing stakeholder views, the more likely they are to accept and act upon the results. Where there are specific clients for an evaluation good communication between them and the evaluators should take place before, during and after the evaluation.

4.6 Education of Evaluators

At present there is a minimum level of training available for evaluators in Europe, particularly those who move into evaluation as part of a government administrative career. There are several longer-term courses in science and technology policy and management that include elements important to research evaluation (including the opportunity to carry out Masters and Doctoral theses on the topic) but specific training only exists as short courses.

In this environment, the importance of networking is emphasised. The European evaluators network and the internal networks within the Commission allow sharing of experiences and lessons. OECD also provides a forum where these exchanges can take place. Such activities are probably sufficient to develop intelligent customers for

evaluation – able for example to specify and monitor evaluation studies. Given the rotation normal in government careers this is probably as much as can be expected.

For this reason it is necessary for evaluation expertise at a deeper level to be developed in universities and specialised consultancy companies. The methods described in the evaluation toolbox generally require a high degree of skill and experience to use them. In some cases it is necessary to have a background in the discipline from which they originate, for example economics or econometrics for micro and macro methods, and social science methods for some of the qualitative approaches. Even where aspects of the skills such as network analysis can be relatively easily acquired, interpreting the results requires substantial experience.

If there is a lesson for Japan here it is the need to build a community of evaluators in a setting more permanent in its personnel than the Ministry. One of the best ways to build such expertise is through sponsorship of research on evaluation – the SPEAR/MONITOR Programme marked a critical phase in building the European Community of evaluators – most of its leaders were significantly involved in that activity at an earlier stage of their careers. More recently projects such as ASIF and EPUB have reinforced and extended such networks, as well as their principal objectives of collating and assessing good practice in evaluation.

5 Conclusion

This paper set out to assess the relevance of the experience of the European Commission in the evaluation of Research and Technological Development, and also of the specific findings from the EPUB Toolbox, for the practice of evaluation in Japan. This exercise cannot be seen entirely as an independent comparison since the relatively late entry of Japan to the practice of evaluation of this type has meant that many of the best elements of foreign practice have already been reviewed and to a certain extent incorporated in to practice in Japan. As with all forms of transfer of policy measures and approaches, some elements are inevitably closely bound to a particular administrative culture and do not transfer directly. It is also true that many examples of the use of relevant techniques could already be found in Japan. Perhaps these have been less systematically applied in an evaluation context, certainly by comparison with individual European countries though not by comparison with the Commission. The latter body has often been an important sponsor of methodological development but such work has remained at the level of “studies” rather than entered the formal evaluation system to challenge and extend the views of panelists. An opportunity exists for Japan to move ahead in this respect by maintaining the rapid progress already made in the past few years.