INNO-Appraisal

Understanding Evaluation of Innovation Policy in Europe

Final Report

February 2010

Manchester Institute of Innovation Research (formerly PREST), MBS, The University of Manchester

Prof. Jakob Edler (Project Coordinator)
Dr Paul Cunningham
Abdullah Gök
Dr John Rigby

Atlantis Consulting

Effie Amanatidou
Ioanna Garefi

ISI-Fraunhofer

Dr Susanne Bührer
Dr Stephanie Daimer

Joanneum Research

Michael Dinges
Dr Martin Berger
Julia Schmidmayer

Wise Guys Ltd.

Ken Guy[1]

---

[1] Wise Guys Ltd. formally left the project on May 15, 2009, though Ken Guy continued to contribute to the project in his capacity as a visiting scientist at JRC-IPTS, Seville.

Table of Contents

Part	Chapter	Title	Page Number
		Executive Summary	i
	1	Introduction	1
I	2	Methodology	7
	3	The Use of Evaluation in Innovation Policy in Europe – A Statistical Analysis	23
II	4	Usefulness of Evaluations	87
	5	The Role of Impact Assessment in Evaluation	125
	6	Exploring the Use of Behavioural Additionality	151
	7	Evaluation in the Context of Structural Funds: Impact on Evaluation Culture and Institutional Build u	201
III	8	Country Report: Austria	243
	9	Country Report: Germany	265
	10	Evaluation in the United Kingdom	283
	11	The Case of the Mediterranean Countries	307
IV	12	Conclusions and Ways Forward	327
Annex	2A	Data Collection Template	Separate Document
	2B	Consolidated Data Collection Statistics	
	2C	Data Collection Manual	
	5A	Data Tables for Chapter 5	
	6A	Data Tables for Chapter 6	
	6B	Case Study Guide for Chapter 6	
	7A	Interviewed Experts List and Interview Template for Chapter 7	

Acknowledgements

This project breaks new ground. For the first time, it takes stock, categorises, and analyses evaluation practice in innovation policy in EU Member States. The project ran from January 2007 to January 2010 and would have not been possible without the help and support of many people.

During all of this time, a steering committee has overseen or followed the work, reflected on its various steps and come together to two key events. The advice given by the members of that committee has been extremely valuable. The project team would like to thank the policy makers Ulrike Blankenfeld (DE), Mark Beatson (UK), Luisa Henriques (ES) Nick Constantopoulos (GR), Rupert Pichler (AT), Jari Romanainien (FI), as well as our colleagues Aleardo Fulrani and Bart Kamp (INNOVA Europe), Phil Shapira (MBS, Manchester) and Anthony Arundel (MERIT) very much for their advice and for their valuable time. Without this advice, the reassurance and constructive critique and the ideas for new avenues the study team would have felt much less secure in pushing ahead with such a novel methodology.

We also would like to thank DG Entreprise for the opportunity to do such rewarding work. An especially heartfelt thank you goes to Alberto Licciardello, who has done a tremendous job as project officer to help to keep it all together and focused, to support in times of need, to coordinate with other services, to inspire the analysis and to disseminate our activities to colleagues inside and outside of the Commission. Finally, the project would like to thank all the participants at the INNO-Appraisal workshop in September 2009; the feedback gained at this event has been tremendously important and encouraging in the final stages of this project. Finally, our thanks goes to the numerous policy makers across Europe who have participated in this study, by filling in templates about the evaluations they have commissioned and by agreeing to be interviewed and otherwise give feedback. Without this openness and responsiveness, the study would simply have been impossible.

On behalf of the INNO-Appraisal team

Jakob Edler

PRO INNO EUROPE
INNO
APPRAISAL

Executive Summary

The **aim** of the INNO-Appraisal project was to contribute to a **better understanding of how evaluation is currently used in innovation policy in Europe**, and **how evaluation contributes to policy making**. INNO-Appraisal was the first systematic attempt to provide an overview of evaluation practice in Europe. By doing so, it sought to achieve a **second, equally important aim**: i.e. to **render evaluation practice accessible** to the policy and evaluation community. A **third aim** was to contribute to a better-informed evaluation discourse across a better-networked evaluation community in Europe.

To achieve these aims, the project spent three years taking stock of and assessing evaluations in the area of innovation policy across Europe. It applied a novel and complex approach, combining qualitative in-depth analysis (case studies) and sophisticated quantitative analysis on the basis of a new form of data collection. The basis for the evaluation report collection was the EU innovation policy database Trendchart over the period 2002 to 2007. The project designed and made use of a web-based template to allow a systematic characterisation of all selected evaluation reports. It then interacted with policy makers in order to verify and amend these characterisations. The template data was then used to conduct a statistical analysis of the whole sample and further analyses of sub-samples relating to specific questions and case studies. To make these evaluation reports accessible, a repository of evaluation reports was created and placed on the INNO-Appraisal webpage. This repository allows interested parties to search for and download evaluation reports. It was also designed to allow a keyword search using the categorisation scheme on which the analytical template for each report was based. Thus, policy makers can now perform specific queries tailored to their specific needs, e.g. searches for examples of the application of particular methods, the coverage of certain topics or the evaluation of similar types of programme.

Descriptions of the approach adopted by the project, its various interim results and the repository itself have also been widely disseminated throughout the PRO—INNO® community and the wider policy and analyst community in innovation policy in Europe.

Thus, the major contributions of the project to the evaluation community in Europe and evaluation discourse in general are:

(1) **An analysis of evaluation practice with some in-depth topic-oriented and country case studies (this report)**,

(2) the **repository** on the INNO-Appraisal webpage, with all its various search and download functionalities and its legacy role as a stockpile of evaluation reports and activities (http://www.proinno-europe.eu/appraisal).

In conjunction with earlier reports, presentations and interactions detailing interim results, this report (and subsequent outputs) and the repository itself should contribute to an **improved policy discourse** in the EU and beyond. It needs to be stressed, however, that the report does **not** constitute another ‘evaluation manual’. Rather, it is analytical in nature and provides a service to the Community by making evaluation in Europe more ‘tangible’.

This executive summary encapsulates the major findings of the analytical part of the project in some detail to reflect the depth and breadth of the analysis and to avoid undue simplifications.

1. The characteristics of evaluations in innovation policy in Europe

1.1. In a nutshell: basic characteristics of evaluation practice in innovation policy in Europe

The INNO-Appraisal database contains evaluation of a whole range of different policy measures that are covered in the Trendchart Database between 2002 and 2007. Reflecting the innovation practice across Europe, its majority of evaluations are concerned with direct financial support for innovation activities, and two thirds of the underlying sample measures in the database are geared towards involving Universities and public research institutes. The database of evaluations covers all European countries, with an interesting bias towards Austria which has an exceptionally high number of innovation policy measures reported in Trendchart and an extensive evaluation activity. The repository, as basis for the analysis, cannot claim to cover all innovation policy evaluation in all countries to the same degree, as countries represent their activities differently in the Trendchart database, the basis of the analysis. For example, Finland and Sweden are underreported in the database as many of their evaluations are undertaken within the programme portfolio of large agencies so that many individual measures are not flagged up in the Trendchart database. To get an understanding of the meaning of country contexts, however, later sections deliver in-depth country cases of Austria, Germany, UK and Mediterranean countries. The repository also covers extensively evaluation of structural fund measures, as slightly more than 20% of all evaluations are performed in the context of structural funds. The **total number of evaluation reports** is **242**, of which **216** could be meaningfully analysed by the project team (and thus used for the statistical analysis presented), and **146** were amended and **verified by policy makers** (used for specific, judgemental and policy related parts of the statistical analysis). The number of publishable evaluation reports in the project repository is 173.[2]

Commissioning and design: Evaluation is found to be an integral part of innovation policy, as roughly **50% of the measures** that are evaluated have a **pre-determined budget** for evaluation and **two thirds** are **foreseen and planned** in the measure design. More than 90% of evaluations are sponsored by the programme owners themselves, only a minority are jointly sponsored with other bodies or entirely external (10%). **Almost half** of the evaluations follow an **open tender procedure**, one fifth are done through closed tender, one fifth are performed by external evaluators without a tender and 15% are done internally. For those evaluations which have **a tender**, a **large majority clearly specified the objectives**, whilst at the same time, two thirds of the tender documents left the choice of methods to the evaluators.

***Timing**:* More than 40 % of the database is interim evaluations. This bias against ex post (30%), however, stems partly from the selection method focusing on live Trendchart policies within a certain period of time. The database contains both formative (33%) and summative (21) evaluations, while **the majority combines both summative and formative aspects**.

Topics: The topics covered in evaluations are obviously broad. In very general terms, **effectiveness** and **consistency** appear to be **slightly more important** than programme efficiency issues, while the in-depth look at **project efficiency** is **much less common** (below 50%). We also find a certain **clustering of topics** covered. Two thirds of all evaluations cover at least one form of additionality

---

[2] For a series of methodological and database specific reasons one cannot give a statistical data as for the share of policy measures that are evaluated within Trendchart in the period covered.

(input, output and behavioural), and many of those evaluations tend to include the project level in order to understand those additionalities. Gender (24%) and minority (7%) are least common. In terms of impact, technological and economical are most important, and environmental impacts (still) least important (28%).

Methodology and data sources: In terms of methodology, we find a whole range of methods applied, however, some general strong trends are obvious. **Descriptive statistics** are the **most common** approach, applied by more than three quarters of all evaluations, while case studies – to understand contexts and developments over time – are performed only by 41%. **More sophisticated,** quantitative approaches are used **even more selectively**, e.g. 23% perform econometric analysis, 17% network analysis. Interestingly, 80% claim to use monitoring data and 70% to use existing surveys and databases as a basis for the analysis. However, it appears that this kind of data is insufficient to be used for specific evaluation questions such as networking or behavioural additionality. The most important *pro-active* **data collection is done through interviews and participant surveys**. **Technometric analysis** in innovation policy plays **no significant role at all** (2%), it appears that for the analysis of technological substance in projects peers are used (20%).

***Quality*:** As for overall quality of evaluations, the database shows very mixed results along nine different quality aspects. For a general picture a simple binary quality index has been constructed. All evaluations that score more than 3 on a Likert scale (1 being very low, 5 being very high) in *each* of *four* selected quality variables are defined as being of high quality. 61% of the evaluations show an overall positive quality index. This means that **almost 40% of the evaluations have serious quality problems in at least one key quality dimensions**. This finding is confirmed through an auto-correlation analysis: Many evaluations are either good in a whole set of quality criteria or perform rather badly across the board.

The policy use of evaluations: While almost all evaluations are targeted towards policy makers and programme management, **only 50**% of the evaluations are **targeted towards the users** of the programme and **less than one third to the general public**. Evaluations are obviously **not extensively used to mobilise** the community, policy makers themselves rate the breadth and depth of actual discussion about evaluation results only as moderate.

Most evaluations, as is to be expected, contain recommendations for policy and programme management, **only a minority of evaluations is purely analytical**. The **usefulness of the recommendations** for various aspects of policy learning and improvement that were tested is **moderate** and appears to have room for improvement. In principle, **evaluations are not linked with major, radical consequences**, those appear to be the result of more general policy considerations. However, they are **important for minor re-**design of measures or their **prolongation and extension**. In 17% of all cases they are also used to improve other or future policy measures.

1.2. Determinants of evaluation practice, quality and consequences

There is a certain degree of convergence of evaluation practice across different policy measure. We find **surprisingly little variation between different policy measures** as regards a **whole range of evaluation characteristics**, such as tender procedures, internal vs. external evaluators, coverage of topics and impacts and even use of some of the data collection approaches and methods and even targeted audiences. It shows that other factors, such as organisational and country specific traditions, topics to be covered and general practices dominate the design and implementation of evaluations to a large extent, not so much the evaluation object – the policy measure – itself.

However, the **type of measure makes *some* difference** as for evaluation design and implementation. Certain specific types of programmes show a specific application of tailored methods and data collection approaches (e.g. network analysis and case study approaches for networking and cluster programmes). We also find variation in the **use and dissemination of evaluation** between policy measures, with - for example – **complex networking programmes targeting beneficiaries much more often** as those measures are complex and need explanation and formation. Furthermore, evaluations for direct financial support measures and cluster, technology transfer and networking measures are more likely to be perceived as being of good quality, while **evaluations for softer measures** such as **management support measures** or **diffusion measures** are of **lower quality**. In addition, there seems to be a **poorly developed evaluation practice for diffusion measures**, which – in addition – do not take societal and environmental impacts into account as broadly as expected, and that are perceived to be of less usefulness to policy makers.

Evaluations are often **influenced by external sponsors** of the policy measures. While they do not impose methods they introduce a **bias towards social and environmental impacts and gender** and minority issues. The external sponsors, it seems, are one major reason behind a **certain grouping** of evaluations around topics that we observe, some being more concerned with quantitative, hard economic and technological outputs, and others interested in social, environmental impacts etc.

Evaluators in innovation policy appear to apply a **form follows function** approach; they tailor their approaches according to the need for topics and impacts to be covered. For example, evaluations interested in strategy development and policy issues more generally also look at consistency and vastly use interviews and other qualitative methods. Evaluations more concerned with **effectiveness** rely on (often simple) **statistical analysis and data**, and the **use of peers**, although limited, is **strongly linked to quality of output**. Those evaluations more concerned with **efficiency and project level issues**, in turn, tend to look for different kinds of **additionality** and rely on surveys, interviews and, less broadly, though, on case studies. Further, formative and ex ante evaluations tend to analyse consistency issues more broadly than other evaluations (i.e. to assess and re-adjust the overall match), and they do so be using slightly more qualitative methods.

A deeper look into the **determinants of quality assessments** reveals that policy makers see **room for improvement** as regards the **coverage of the broader context**, the **application of advanced quantitative** and some qualitative methods and the **documentation of information sources**. In contrast, evaluations covering **technological and scientific impact** and those using survey methods and peer review are perceived as being of a higher quality. Summative evaluations appear to be perceived as being of higher quality than formative evaluations, and indeed they are more widely discussed within government than formative ones. Formative evaluations, it seems, are a tool for improvement for the programme owners and beneficiaries, while the messages of summative evaluations are used for wider discourse and justification.

Interestingly, **quality** does not differ between evaluations that are undertaken by external evaluators and those performed internally. Equally, evaluations are **not** perceived to be of **higher quality** if they are **in-built in policy measures from the start** and **have a dedicated budget** within the policy measure. However, one important finding is that quality is lower for evaluations that are commissioned by external sponsors or policy bodies. In contrast, open tenders yield evaluations of better quality.

Quality, finally, **makes a difference** when it comes to the **dissemination and exploitation** of evaluations. The better an evaluation, the more likely it is discussed within and outside government. In addition, evaluations that ex ante are targeted to the wider public and policy analysts (and not only to the programme management) are also of higher quality.

The analysis also revealed that evaluations have **a limited set of consequences. Radical consequences** (termination of programmes) are **very rarely a result of an evaluation**, but rather they appear to be the consequence of principle policy decisions. In contrast **evaluations lead to minor re-design of measures or learning for other measures** and, most often, to **prolongation and extension**. The latter is highly correlated with simple methods, it thus appears that **clarity and simplicity** in the data and methods is part of a **confirmation and incremental approval exercise**. In addition, those evaluations which are intensively discussed within and outside government are those that are more likely to lead to consequences. Finally, quality also is important for evaluation consequences; evaluations of higher quality more often tend to lead to consequences (especially prolongation). The quality aspects most strongly linked to the likelihood for consequences out of evaluations are the extent to which evaluation methods satisfy the Terms of Reference and the purpose of the evaluation.

In a final analytical step the general statistical analysis explored **clusters of evaluations**. Two clusters emerged. One cluster of evaluations is more populated by ex ante evaluations and is concerned with programme efficiency issues and, by nature, more often based on qualitative methods. The second cluster appears to be more ex post and interim, being broader in its coverage and more concerned with different forms of outcome/impact, thereby mobilising more quantitative approaches and oriented towards the policy community rather than the beneficiaries. This cluster of evaluations is more often used for decision about prolongation or re-design of measures.

2. In-depth analysis of selected evaluation issues

Four themes of evaluation have been identified as being of specific importance to stakeholders and the evaluation community have been studied in considerable depth; usefulness, measurement of impact, behavioural additionality and structural fund evaluation.

2.1. Usefulness of evaluation

The analysis of the usefulness (or utility) of evaluations sets the broader context of policy interventions within a policy mix, and the accompanying need for policy makers to be able to judge the effectiveness and efficiency of their interventions through the use of a range of governance tools, including appraisal, monitoring and evaluation. It is clear, from the policy mix concept, that the information gained from these tools should not be restricted to the subject of the assessment, but should also be relevant to the design and operation of contemporaneous or subsequent policy instruments: such requirements define the issue of usefulness.

The report then discusses what is meant by usefulness and utility in the context of the evaluation of innovation support measures. **Three major purposes** for evaluation are identified: operational learning, policy feedback and system impact. Overall, it is suggested that, to be useful, evaluations must provide information on: the effectiveness of design, the effectiveness of management, the effectiveness of implementation, the effectiveness of the evaluation itself, the achievement of objectives, and the broader impacts of the instrument. However, it is recognised that usefulness may also be impacted by other factors such as audiences and sponsor demands.

A number of factors are then examined whereby the **utility** of **evaluations may be increased**. Possible routes include **increasing the rigour** (and hence 'quality') of an evaluation, obtaining the **compliance and trust of stakeholders**, improving the **transparency of methodologies** (assuming an informed audience of policy makers is present), and the use of **clear and measurable objectives**. The incorporation of evaluation into the **overall policy cycle** is seen as a clear route to improving the usefulness of its outcomes.

The chapter next deals with the approaches employed in the analysis of the survey results to determine the extent of usefulness of the evaluations reviewed. Two lines of analysis were followed: looking for evidence of utility provided by the responses and testing of hypothesised links between utility and other database variables. As the questionnaire did not specifically pose a direct question on the usefulness of the evaluation (which would have prompted highly subjective responses unsuitable for quantitative analysis), it was necessary to develop a proxy for usefulness based on the extent to which the evaluation report's recommendations had been useful (a point addressed by specific questions in the questionnaire template). This **proxy indicator** (for overall usefulness) could be broken down into **internal utility** (relating to changes to the programme under appraisal) and **external utility** (relating to changes to contemporaneous or subsequent programmes).

The analysis then examined a number of the **database variables** for **links with usefulness**. The main points to emerge were:

- 84% of evaluations examined had contained recommendations, with an almost equal balance between internal recommendations (relevant to the subject programme) and external recommendations (relevant to future programmes or to broader policy formulation).
- Evaluations addressing internal aspects of the programme had a slightly higher usefulness than those addressing external aspects.
- **Significant positive correlations with** at least one aspect of **usefulness** were identified for:
    - The use of an open tendering process when commissioning and evaluation
    - The use of external evaluators
    - The timing of the evaluation (*ex ante*, interim, *ex post*, etc.)
    - Summative over formative evaluations
    - Non-Structural Fund evaluations (i.e. a negative correlation between Structural Fund evaluations and utility)
    - Non-portfolio type evaluations (i.e. a negative correlation between portfolio type evaluations and utility)
    - Non-conditional evaluations (i.e. a negative correlation between conditional evaluations and utility)
    - Evaluations that examined the topics of goal attainment and effectiveness and policy/ strategy development

  - Evaluations that examined scientific impact and technological impact on the participants and beyond
  - Evaluations that employed case study analysis; participant surveys; interviews; focus groups/workshops and meetings; peer review
  - Evaluations that resulted in a minor redesign or expansion/prolongation of the measure
  - Evaluations sponsored by programme managers, other government departments or other public bodies
  - Evaluations not conducted primarily for auditors/financial authorities
  - Evaluations whose reports were published in English
  - Certain dimensions relating to the quality of the evaluation
- **Negative correlations** with at least one aspect of usefulness were observed for:
  - Evaluations that examined input additionality and environmental impacts
  - Evaluations that employed input/output analyses; context analysis; group comparison approaches; cost/benefit approaches; existing surveys and databases
- **No significant correlations** with any aspect of usefulness were detected for:
  - Evaluations planned during the design of the measure
  - Presence of a dedicated budget for the evaluation
  - Evaluations conducted primarily for policymakers (government officials) and programme management
  - Evaluations that examined outputs, outcomes and impacts; quality of outputs; value for money; programme/project implementation efficiency
  - Evaluations that employed monitoring data
  - Evaluations that had wider levels of availability
  - Evaluations where a major redesign of the measure resulted
- External utility was more highly rated in Germany and the Netherlands, whilst internal usefulness was more highly rated in Greece, Sweden and the UK
- The evaluations of **measures for science-industry cooperation** were **significantly more useful** across all categories of usefulness. Evaluations of measures aimed at the creation of start-ups and spin-offs were also significantly useful (external and overall).

Whilst a number of the statistically significant associations between usefulness and the survey variables were anticipated, it is harder to explain some of the negative correlations or where no correlations were detected. Several of the latter might be explained by the relatively low number of cases available within the analysis, whilst the prevalence of Structural Fund evaluations within the sample could also provide an explanation.

In conclusion, the results of the analyses present a **mixed picture**, confirming some expectations yet failing to confirm or even refuting other expectations. As with most research endeavours, it is clear that further investigations are required into the aspect of usefulness and it is hoped that this study offers a valuable starting point. Nevertheless, the results do tend to support the overall conclusion (which is also based on the direct input of policymakers in the field) that **usefulness is a highly subjective and context specific issue** and that, as a broad rule of thumb, an evaluation may be **considered useful if it delivers the Terms of Reference** in a **consistent manner** and if it provides **actionable recommendations** and delivers **evidence as for value for money**. Usefulness can be defined as the degree to which there is feedback on policy and if the evaluation process delivered some degree of policy learning.

2.2. Measuring impact

While there is extensive academic debate about an ideal-type setting for impact measurement in evaluations, it is a quite different matter how impact assessments are performed in reality. Most often, impact assessment is rather limited and simplistic in its approach within the reality of service contracts for programme owners, most probably including budget restrictions, specific "customer" needs, and tough schedules. One in-depth analysis of the INNO-Appraisal database has looked systematically at the application of impact measurement in evaluation of innovation policy. It explored if there is some systematic use of methods, i.e. whether there are certain sets of methods which are employed for specific policy measures and in specific contexts, and whether, and also to what extent evaluation studies of policy programmes have an impact on future innovation policy.

In sum, the quantitative analysis of the database shows a number of interesting results: In general,

- Impact assessment is a central function of evaluation studies: A large number of studies across Europe claim to do impact assessment, currently most important are economic impacts.
- Impact assessments appear to be central and wide-spread across Europe
- Impact studies of structural fund evaluations differ significantly from impact studies of national innovation programmes.

As for specific types of impact, we find that

- Typically the use of a very **broad definition** of impact assessment, including all types of effects
- Assessment of economic impact is most dominant, other impact types of importance are technological and societal impacts (not scientific and environmental impacts)
- The assessment of **new impact types** (apart from economic or technological) is still rather uncommon. Societal impacts are often covered with an estimation of new jobs having been created, but other topics, such as gender impacts are quite rare.
- A high number of evaluations claim to assess **indirect impacts**, i.e. spill-over effects beyond the participants of a programme. This is, given the methodological difficulties for assessing economic or societal impacts, a surprising result. This result seems to reflect the demand for results on these spill-over effects.
- **Additionality** concepts are well established beyond the UK. They are employed by half of the evaluations in the sample. This is also true for behavioral additionality which has obviously become an integral part of the idea of additionality.
- Structural fund evaluations more often cover social and environmental impacts.

Methods used

- Almost the whole toolbox of possible **methods** is employed for impact assessment, including elaborate methods such as a control group approach.
- Most of the impact assessments are qualitative and part of broader evaluation studies.

- There are only few quantitative impact assessments using elaborated quasi-experimental designs like control-group approaches.
- Impact assessment is typically not a mere econometric exercise, but **often used in a** contextually **sensitive way.**

Policy Cycle

- Impact assessment is not a clearly retrospective element of evaluation. Often, it is also used in the form of ex-ante impact assessment and in accompanying evaluations.
- Evaluations which include impact assessments, in particular the assessment of societal impacts, are more often used for external communication. Experts confirm that impact assessment is in particular important for legitimizing the political interventions.
- If impact assessments are included into evaluations this leads to higher quality scores.
- With respect to usefulness, evaluations of (single) national programmes seem to be more useful for policy makers than structural fund evaluations.

The analysis and the interviews indicate a **set of clear recommendations**. Most important issues from the perspective of the policy maker are:

1) Impact assessments are an important part of evaluations, but **should not be isolated**. Ideally, impact assessment is **integrated into a broader, more holistic evaluation framework** (e.g. covering context analysis systematically), only then can it fully be understood.
2) Evaluators have responded to the demand for quantitative results and employ a variety of (elaborate) methods to achieve them. However, in most cases it seems that **the combination of qualitative and quantitative analysis** can cope more adequately with impact assessments, as many impacts are not quantifiable at all.
3) Many pitfalls of impact studies can be avoided by a **constant communication between policy makers and evaluators** during the process of evaluation. This leads to transparency for the whole evaluation process in order to realize learning and to cope with methodological challenges.
4) As impact assessments clearly pursue the two purposes of learning and legitimation, two types of recommendations might be considered as a result of impact evaluations: those designed **for policy improvement** implemented by the programme owners / managers and those **directed to higher levels**, which serve the legitimation aspect.

For the future, it is useful to consider further impact dimensions **to a greater extent** than in the past. Additionally, as we expect more mission oriented policy programmes where other topics like sustainability, customer needs and the structural / regional development might become more important. Thus, impact assessment will have to be **broadened considerably in the future**. Further, with examination of the demographic challenges and shortages of a skilled workforce in most European countries, the issue to integrate larger parts of society to the research sector will become even more relevant than in the past and therefore impact assessments should also address **gender and minority issues** in more detail. Finally, the still prominent aspect of Behavioural Additionality in Innovation Programmes (e.g. innovation management, risk aversion) will remain important.

For impact assessment, this all means that it will become **even more demanding** to measure the intended effects – at least quantitatively. Given that non-economic impacts will gain more and more in importance this would mean that new sets of criteria and indicators will have to be defined, and most likely many of these indicators will be of a qualitative nature. More public support for **experimental evaluation designs** (including meta-evaluations at national as well as European level) could help to identify the most promising ways to identify new impact types.

However, given the multiplication of goals and increasing pressure as to economic effects impact assessment, the requirement to establish the relevance and rank different impact dimensions will become all the more necessary. Equally, policy must reflect if, and to what extent the large set of impact dimensions can really be achieved by one single measure instrument. The programme objectives have to correspond with an appropriate mix of policy instruments and have the right balance between direct and indirect funding. Additionally, policy design has to be very aware about the prerequisites for (behavioural and system) change which cannot entirely be influenced by singular measures. By definition, impact assessment can only be one, even if essential, part of evaluation to support those policies.

2.3. Behavioural additionality

Behavioural Additionality is still a rather novel, but already a **key topic for evaluations**. The concept has **enlarged** our thinking about the **effects of innovation policy** to include, more systematically, **learning** as a key outcome in itself, enabling further, and broader, and more sustainable innovation. Behavioural evaluation is a case of reflexive intelligence, whereby working on understanding the concept and applying it to innovation policy itself co-evolves with innovation policy concepts that take learning into account much more profoundly. Evaluation practice and conceptualisation on the one hand and innovation policy development on the other hand have re-enforced each other. The empirical analysis on behavioural additionality in this study rests on three pillars, a statistical analysis, a text analysis of evaluation reports and a set of interview based case studies of evaluations.

In the **academic literature**, the term is understood in at least **four different conceptualisations of behavioural additionality**, namely, i) as an extension of input additionality, ii) as change in the non-persistent behaviour related to R&D and innovation activities, iii) as change in the persistent behaviour related to R&D and innovation activities and iv) as change in the general conduct of the firm with substantial reference to the building blocks of behaviour.

Against this background, a text analysis of 33 selected evaluations demonstrated that the diversity of understandings is reflected as well in **evaluation practice**, where we also find **at least four** different **understandings** of the concept. They are distributed rather evenly in the sample – and thus there is yet no dominant understanding established just like the case in the scholarly literature. These understandings differ in the their conceptual outreach, ranging from collaboration (non-persistent) in R&D and innovation only – the most narrow type – to persistent change in management practices more broadly, beyond R&D and innovation – the broadest type. The types overlap, but not entirely match four ideal types as defined in the vast literature on the concept.

The analysis of the INNO-Appraisal database aims to show if and how evaluations differ that apply the concept from those that do not. For the first time this allows us to get a systematic picture of the nature of behavioural additionality in practice. The core results are as follows.

The data analysis shows that behavioural additionality is a well established concept in evaluations, **50% of all reports in the database employ it**, explicitly or implicitly. The concept is more often used for networking and technology transfer concepts, which is consistent with the need for learning, networking and cooperation in those programmes. The behavioural additionality concept is most **often** used **in conjunction with input and/or output additionality**. It appears to be more important in evaluations that are also concerned with **project level evaluations**, not only programme level, which again is consistent with the basic idea of understanding the micro level in order to understand the macro effect. The concept is **less common in portfolio and structural fund evaluations** as those often do not look at the project level.

While there is no difference between evaluations that are sponsored by the programme owners themselves or by other bodies, we observe that the concept is slightly less often applied in internal evaluations. The application **needs specific expertise** and in-depth qualitative approaches which seem to be best conducted by external evaluators. However, this does not imply that evaluators are more keen to apply it than policy makers, since the concept is more often applied in those evaluations in the database that specify the methodology in the terms of reference – and thus express **a clear demand for behavioural additionality** approaches. Our in-depth case studies indeed confirm that both evaluators and policy makers can be the source for the application of the concept, it is not entirely evaluator driven.

Interestingly, and neglecting its full potential, behavioural additionality is not as common in accompanying evaluations as one would assume given the focus on interaction and learning and the need to re-adjust programme and implementation should learning effects not be observed in real time. The concept is used in formative evaluations, but not as extensively as one would think. Similarly, evaluations that cover behavioural additionality are less likely to look at social and environmental impact, but **much more at scientific and technological impact** than the whole sample, while the concept is equally concerned with economic impact than the whole sample of evaluations.

As for methods, behavioural additionality evaluations are **more qualitative** and apply those methods with greater quality, however, the extent of case study analysis is not as broad as one would expect. Behavioural additionality evaluations also **use surveys more often**, while they **cannot rely on existing data or monitoring data**, pointing towards a need for adjusted monitoring.

Behavioural additionality evaluations are **broadly discussed** across government and beyond government, and they are **more often targeted towards the general public and towards users**. All this points to the **learning and mobilisation** potential of the concept. However, evaluations applying behavioural additionality are not perceived to be significantly more useful for changes in policies than other evaluations (although they perform slightly better in this regard). In terms of concrete consequences of the evaluations that apply behavioural additionality, the major difference to the general dataset is that the former lead **significantly more often to the extension of existing measures**. This again points to the underlying understanding of long term effects and the need for time in programmes that rely on the learning of actors.

The case studies finally confirm the variety of understandings and different application of the concept and the challenges the application of the concept faces. This is true both at the receiving end, with the programme owners, and at the performing end, with the analysts. The cases show that evaluators and policy makers alike are **keen on understanding changes in behaviour better**, but they also confirm that **policy makers strongly demand a demonstration as to how the behavioural change translates into the intended innovation effect**. However, many variables influence change in innovation, attribution remains a **constant challenge** and innovation effects often take considerable time to realise. Evaluations thus must clearly **demonstrate the conceptual link** between behavioural change and the innovation effect. They then must empirically grasp the change in behaviour and try to find robust indications that the conceptual link to innovation effects exists.

As yet, the applied **methodologies** do most often **not fully capture behavioural additionality**. The cases however show that it is possible to differentiate behavioural additionality and define building blocks of behaviour as well as chain of effects. This can be done in a **mix of deductive and** inductive approaches, with a **focus on interaction with the beneficiaries**. But there also is a **delicate balance** between exploring the concept to its full potential through all sorts of differentiation and methodologies on the one hand and **pragmatic considerations** and limits of absorptive capacity on the other hand. Thus, **more experiments with sophisticated methodologies** are called for. Those experiments should then enable us to define sets of meaningful, simplified methodologies that are more effective and efficient than the existing approaches, but do not overburden the process. To that end, there seems to be a huge potential in improving monitoring of programmes to use it for evaluations much more thoroughly.

Finally, the complexity of behavioural additionality **asks for a strong interaction and communication** between those commissioning the evaluation and the evaluators, since key concepts as to the link of behaviour changes to innovation must be shared between them and expectations clarified early on. Sophisticated methods alone do not guarantee the full benefits of the concept, their applications and the results **must be intensively discussed among all stakeholders involved**.

2.4. Evaluation in structural funds

The aim of this focus study is to examine if, and in what ways, the Structural Funds (SF) requirements and regulations related to evaluation influence the evaluation culture, institutional build up and good practice in evaluation in certain countries. It draws upon the results of the questionnaire survey carried as well as the examination of the uptake of SF regulations in three countries, Greece (a Southern European country) and two new Member States, Poland and Malta. The specific countries are examined as indicative examples of how SF evaluation related regulations and provisions are implemented and affect evaluation practices in their specific contexts.

The case study collection and analysis of data, information and stakeholder views is guided by the **following hypotheses**:

- SF requirements may lead to specific characteristics in delivery and practice of evaluation
- SF requirements may lead to higher quality evaluations
- High quality SF evaluations may have greater impact
- SF regulations demand high standards on structures and processes that inevitably need some institutional learning and structure building

SF regulations do seem to **lead to specific characteristics** in the delivery and practice of evaluation. They tend to be **built in to the design phase of a programme/measure** as they are a requirement in the SF implementation. They usually also meet the requirement to make the **results publicly available** through publication of the evaluation report. Recommendations mainly relate to the programme / measure being appraised in terms of design, management and implementation clearly **reflecting the orientation of the SF evaluations**.

SF requirements also seem to contribute to guiding the evaluation topics covered under the different evaluation types (ex-ante, interim, ex-post) as well as the data analysis methods used (but not the data collection methods). Yet, SF guidelines seem to more or less repeat what is suggested by international practice in evaluation and thus is also followed by non SF type evaluations. This might be the reason why no major differences exist when studying the results within the same evaluation type (ex-ante, or ex-post for example) across the two populations (SF and non SF).

SF requirements do **not** seem to lead to **higher quality appraisals** and **even high quality SF evaluations** do **not** lead to **high impact in terms of usefulness of recommendations and dissemination of results**. However, the suggestion to use **independent (external) evaluators** does seem to contribute to **higher quality** SF evaluations.

The country cases provide possible **explanations** for the survey results. The **fragmentation** among the key actors in the national innovation system in Greece, for example, and the fact that there is only typical abidance to SF regulations can explain why the results of SF evaluations are discussed with government and wider stakeholders only to a limited degree.

Abidance by the ‘letter rather than the essence of the law’, in combination with doubts about the suitability of the SF regulations leading to high impact evaluations can explain the limited usefulness of recommendations, as well as the fact that even high quality SF evaluations may not lead to high impacts in terms of usefulness and dissemination of results. The fact that SF regulations and quality standards are only suggested rather than imposed may explain why suggested quality criteria may not be applied in practice.

Finally the country cases show that while SF regulations have caused positive impacts in terms of capacity and structure building, they still fall short in improving institutional learning and establishing sound evaluation systems in the countries examined.

3. Country level analysis

3.1. Austria

Having been a laggard in terms of RTI (Research, Technology and Innovation) investments until the mid-nineties, both public and private entities have increased R&D investment efforts tremendously in the last decade. Austria has exceeded the average R&D intensity level of the EU-15 and the OECD countries. But not only RTI funding has increased: Austria has a large stock of innovation promotion measures at hand: Apart from generous bottom-up RTI project funding schemes, a remarkable number of thematic R&D programmes, structural programmes, and tax incentives exist. Despite good overall conditions there are a series of systemic challenges that still need to be addressed (e.g. poor performance of the Austrian higher education system, insufficient framework conditions as regards regulations, poor private and public funding for innovative start-ups and spin offs).

During the catching-up process, RTI programmes were the most preferred way to address policy challenges. In this time, the use of evaluations increased dramatically. Evidence for the increased relevance of innovation policy evaluation is provided not only by evaluation counts, but by changes in the legal conditions for evaluations, measures to foster an evaluation culture, the transparency of evaluation results, and the high number of evaluation activities.

With 34 appraisal reports, Austria has the highest share of innovation appraisals in the INNO-Appraisal database. Some distinct features of these evaluations are presented.

The majority of appraisals are carried out mid-term during one point in the programme's lifetime. Mainly, a supportive purpose is followed as policy makers respectively programme managers' needs advise how to enhance programme implementation. Only a limited number of topics are addressed: Appraisals focus mainly on policy/ strategy development, output counts, and consistency matters. Whereas behavioural additionality issues are rather prominent in Austria, input and output additionality issues as well as quality of outputs are only considered in a limited number of evaluations. Technological, economic, and socio-economic impact dimensions are missing by large, or only refer to programme participants.

Low cost data gathering and data analysis methods prevail (descriptive statistics, context analyses, interviews, and monitoring data). Most commonly a mixed methodological approach where quantitative and qualitative methodologies are combined is used.

Compared with the other countries in the dataset, we see a significantly lower coverage of input and output additionality issues, also the quality of outputs is widely neglected. Only a limited number of Austrian appraisals deal with impact at all: For every impact dimension, coverage is lower in Austria than in the other countries of the dataset. If impact dimensions are covered they focus on direct impact rather than on the participants and beyond.

Partly, the low coverage of impact dimensions and certain topics might be due to the formative purpose of most evaluations. Another reason for the discrepancies is the high coverage of Austrian appraisals in the database. Whereas in Austria almost the full range of appraisals conducted in the field of innovation policy is covered, it is more likely that only bigger evaluations are covered in the other countries; significant differences as regards the tender procedure point in this direction.

Despite the intermingled picture as regards evaluation topics used, the quality of evaluations is perceived to be high by respondents. Given the evaluation purpose, the methods used tend to be considered as appropriate. Especially, recommendations concerning changes to the management and implementation of RTI programmes were perceived to be useful. Forward-looking advice was regarded as helpful for the design and implementation of future policy measures.

Nevertheless, due to the high number of evaluative activities, an increasing evaluation fatigue can be witnessed. Criticism was raised, that mechanisms ensuring that the results of evaluations are fed back into policy formulation and implementation are missing. In this respect, more thoughts need to be spent on the concrete purpose of planned evaluation activities, and the role of evaluations for policy implementation.

3.2. Germany

Four major findings for innovation policies and the evaluation practice make Germany an interesting case to study and allow us to draw some recommendations on good evaluation practice.

First, innovation policies in Germany are focussing on high technologies, SMEs and the remaining special situation of the Eastern federal states. This is clearly reflected in the evaluated policy measures in the Inno Appraisal database.

Second, the institutional setup at the federal level provides for quite a systematic approach to evaluation. Almost all programmes are being evaluated. In particular the Ministry of Economics regularly foresees evaluations, when planning new programmes. Open tender procedures and the commission of evaluations to external evaluators are standard. This practice is not only clearly visible in the database. The INNO-Appraisal data shows that this practice leads to particularly high quality of evaluations; specifically, the application of open tender procedures is linked to high quality scores.

Third, evaluation reports are very often publicly available and there is particular interest in the evaluation community. The foundation of the Society of Evaluation and several attempts of standardization have intensified scholarly debates. In fact, there is some standardization of approaches visible, but more importantly, this convergent development takes place at a high quality level and includes the openness of evaluators (and commissioners of evaluations) towards new methods.

Finally, we have evidence from the data as well as from expert interviews that learning is a purpose of the commission of an evaluation. There are many formative evaluations, methods such as focus groups or workshops are often employed, and the results of an evaluation are intensively discussed within government. Generally, it seems that learning does takes place. However, although we find a high number of accompanying (and interim) evaluations in Germany, it seems that learning applies in fewer cases to the evaluated measures themselves, but takes place on a more general level, namely, the overall policy learning for future policy making and programme design. One of the reasons for this is that the aspect of "policy/strategy development" is an integral part of formative evaluations in Germany.

3.3. UK

It is a widely accepted belief, supported by documented evidence, that the UK has a strong culture of evaluation in RTDI policy making. This case study examines the broader context within which the processes of review, assessment, appraisal, monitoring and evaluation are employed within the UK system of innovation policy governance, a system which, due to the broad definition of innovation held in the UK, encompasses a number of policy domains and actors.

In particular, a number of relevant features of the UK innovation policy governance system are considered, including:

- The use of strategic review processes (and a framework for performance monitoring)
- The presence of multiple actors and stakeholders
- Multi-level governance
- The evolutionary shift from direct support to framework support.

The study then looks at the underlying factors and developments that have shaped the evolution of the current system of evaluation practice in innovation policy governance. These are: a) the development of systematic approach to evaluation in the 1970s and 1980s; b) the accumulation of evaluation expertise through limited meta-evaluation that has led to an innovation culture in government which recognises the value of a practical business oriented approach to policy; c) the growing consensus around the neoclassical model of the economy and society; and d) the extension of evaluation activities throughout government as the devolution of policy and programme and project design and their evaluation has been pushed downwards and outwards from Whitehall to the regions.

Current evaluation practices and tools are then reviewed, in the context of recent structural changes in the machinery of governance in the UK, with a focus on those employed by the Department for Trade and Industry (DTI) and its more recent incarnations, the Department for Innovation, Universities and Skills (DIUS) and now the Department for Business, Innovation and Skills (BIS). The overarching influence of HM Treasury across all policy domains (and the imperative of demonstrating 'value for money' from policy intervention) is exemplified by the guiding principles set out in its 'Green Book', whilst the promotion of a systematic approach to the policy cycle and to performance measurement (including the use of appraisal, monitoring and evaluation) is underlined by the use of tools such as business cases, programme plans, balanced scorecards and the ROAME-F tool. Evidence is also provided for the cascading down of this guidance to the regional level of governance.

There is also support for the fact that policy interest in the UK extends beyond the mundane and routine application of evaluation as a formalised requirement and into the more exploratory and learning-oriented application of evaluation as an evolving policy tool which is adaptable to a variety of new and changing contexts. This is evinced by the 'Magenta Book', which provides guidance on social research methods for policy evaluation and endeavours to develop a greater understanding of the use and applicability of various approaches to evaluation, from the broad to the specific level.

Overall, it is clear that there is an extensive literature and a range of embedded practices relating to appraisal and evaluation in the UK policy system, all of which reinforces the view that the country possesses a well developed evaluation culture.

The study ends with a more detailed examination, in the UK context, of a number of issues which the INNO Appraisal survey of evaluation reports sought to investigate. These were:

- The rationale and purpose for an evaluation: primarily this is aimed at ensuring value for money, coupled with policy learning considerations, which can include identifying unanticipated outcomes and spill-over effects.
- The sourcing and selection of evaluators: all evaluators are external, ensuring independence and evaluation competence, with open tendering a preferred option. Evaluators must meet stringent criteria.
- The use of terms of reference and opportunities for innovative evaluation approaches: Terms of reference are set according to established principles; exploratory approaches are encouraged, provided the principal requirements for the evaluation are met.

- The timing of evaluations: depends on context – the rolling nature of UK programmes tends to favour interim evaluation. Monitoring and appraisal are also standard practices.
- The conditionality of evaluations: evaluation is a pre-condition of HM treasury funding for interventions above a certain funding level.
- The use of dedicated budgets for evaluation: Evaluations are always foreseen and budgeted for.
- Planning of evaluations: All programme formulation includes appraisal, monitoring and evaluation as anticipated elements.
- Topics, data collection methods and data analysis methods: these are all highly dependent upon the context and purpose of the innovation support measure under evaluation. The Magenta Book offers guidance on the appropriate methodologies for use.
- Programme impacts: Evaluations tend to look for both anticipated and unanticipated impacts. Again, the Magenta Book provides guidance on programme impact and how it may be measured.
- Sponsors, audiences and the availability of results: Programme managers form the immediate audience although HM Treasury is the ultimate audience and sponsor. Evaluation in BIS is also under scrutiny from a high level steering group. As a rule, all evaluation reports are made publicly available, except in certain cases where confidentiality concerns arise.
- The production and uptake of recommendations: Recommendations, provided they are realistic and economically feasible are generally acted upon. Similarly, they will be published provided confidentiality concerns do not arise.
- Quality and utility: Quality is defined as being fit for purpose, meeting the Terms of Reference and delivering within budget. Quality is an asymptotic function: there is a minimum level of quality that must be achieved for the delivery of the evaluation's objectives. An evaluation is deemed to be useful if the evaluation delivers the Terms of Reference in a consistent manner and if it provides actionable recommendations and delivers value for money

In conclusion, it is clear that the UK does indeed possess an extensive and historically well-developed culture of evaluation which though formalised and set firmly in a framework geared towards the assessment of performance measurement, policy relevance and value for money, is nonetheless adaptable, context sensitive and reflexive and, moreover, practised by a policy community that appreciates it as a key tool for policy learning and improvement.

3.4. The case of the Mediterranean Countries (Cyprus, Greece, Italy, Malta, Portugal and Spain)

The aim of this case study is to examine the present situation in the six Mediterranean countries (Cyprus, Greece, Italy, Malta, Portugal and Spain) with regards to the ways evaluations are carried out. It is mainly based on the results of the specific questionnaire survey carried out under the INNO-Appraisal study and more specifically focuses on the evaluation topics covered, the identified data analysis and collection methods, as well as the level of quality and usefulness of the evaluations. These findings are then compared to the overall results of the INNO-Appraisal study in order to examine possible identified inconsistencies and differences.

Given that the evaluations in the countries under the focus of this study are mainly carried out according to Structural Funds requirements, the results are similar to those of the Structural Funds

(SF) type evaluations examined in the in-depth case study on SF. However, the evidence base is different; all SF type evaluations in the SF case study compared with non SF type evaluations vs. the six countries' results compared with the total results of the INNO-Appraisal survey.

The initial research hypotheses were as follows:

- Specific evaluation topics are covered in the countries examined vs. the overall results;
- Specific data analysis and collection methods are followed in these countries;
- Specific audiences are addressed;
- Specific quality characteristics are covered;
- Specific issues of usefulness, dissemination and consequences are addressed.

The specific case study mainly draws upon the results of the specific questionnaire template survey carried out under the INNO-Appraisal in comparison with the overall results of the project in order to discover differences and draw substantial conclusions, as well as test whether the aforementioned hypotheses made are indeed the case in the Mediterranean group of countries.

The survey has indicated a small number of differences, but mainly across the different evaluation types, rather than across the Mediterranean countries and the overall population. This suggests that what really makes the difference is SF regulations in terms of how the evaluation types are conducted, but do not seem to suggest anything different from what is usually dictated by international practice, something which is also reflected in the overall results. In terms of quality characteristics, all of them are less satisfactory in the case of the Mediterranean countries in comparison with the overall results. Yet, when examining the results in the six countries in isolation, it is interesting to note that almost all quality characteristics score between 3 and 4 on a 1-5 point scale in terms of satisfaction. This fact can be considered a relatively positive impact of SF regulations given the lack of evaluation tradition in these countries. However, despite the relatively good quality of these evaluations, their results are rarely discussed with government cycles or relevant stakeholders, which is another striking difference with the overall results.

4. Conclusions and Ways forward

This study has, for the first time, provided the policy community and the evaluation community in Europe with a statistical account and analysis of evaluation practice in Europe. Evaluation practice in Europe is highly diverse: it differs between countries and it shows an enormous range in terms of methodological approaches, coverage of topics, quality and usefulness. Different institutional settings and policy traditions in countries influence evaluation practice – and vice-versa, as especially the Austrian case has shown. Evaluation has spread across Europe as the structural fund provisions have pushed countries towards evaluation – though with mixed results to date. The analysis presented in this report constitutes an important step forward in our understanding of evaluation. One key consequence, or so the authors of the study hope, is that the results will allow both policy makers and evaluators to reflect on their own practice, on their approach to evaluation and, ultimately, on the use of evaluation.

While readers may draw their own conclusions as to the lessons to be learned from the analyses presented in this report, and while each of the chapters delivers specific insights from which lessons can be drawn, there are a set of key observations that should support further improvements in evaluation practice across Europe. Once a rarity, evaluations are becoming increasingly commonplace, yet the analysis has shown that this does not automatically lead to good quality evaluations and productive learning as a consequence of evaluations. Greater care needs to be taken along the whole policy cycle to ensure that evaluations are correctly designed, implemented and utilised, with close interaction at all stages between those commissioning and those performing the evaluations. Policy makers need to be 'intelligent costumers', they need to have the absorptive capacity to understand what evaluations can deliver and what they cannot deliver. Evaluators, in turn, must ensure quality throughout the process, especially, though not exclusively, in the application of methods and the development of a thorough understanding of the wider policy and political context in which measures are situated.

Further, conditions and practices concerning the discussion of evaluations within government and beyond must be improved. More thought needs to be given at the planning stage to this phase of the process and to the channels of communication that can be exploited, but evaluators themselves also have to bear in mind that the likelihood and quality of subsequent discussions are highly dependent upon the perceived quality of their reports and the clarity with which methodologies are described and results presented. All this then leads to a more fruitful discussion within and across government and better-informed decisions. In future, however, there will be a need for even greater conceptual clarity given the increasing complexity and sophistication of both innovation policy and the evaluation tools needed to assess the impacts of these developments. The case study of behavioural additionality demonstrated how complex it is to turn one important idea into an operational concept that is both theoretically sound and offers added value to policy makers.

Other operational improvements are also needed. These include the more tailored and conscious design and use of monitoring systems, with evaluations building on the data they produce and monitoring becoming an integral part of the learning process. Evaluation, moreover, should be perceived as a mobilising tool for innovation policy at large, a function highly underused.

Finally, a dilemma confronting evaluation has to be noted. In order to provide the new methods and concepts needed to better inform policy, evaluation itself has to be innovative. Yet the commissioners of evaluations are often very conservative, specifying conventional methodological approaches in their terms of reference despite known limitations and shying away from more experimental approaches. Opportunities to push the boundaries of evaluation theory and practice are thus often constrained.

Allowing for more experimentation, however, will become more important in the future. Evaluation practice in Europe will have to follow the principle of 'form follows function' much more closely. The evaluation of innovation policy will have to adapt to new trends in innovation policy and the demands being placed upon it. The analyses in this report have shown a considerable degree of uniformity of evaluation designs across policy measures. Evaluation practice, to a large degree, is an exercise in 'copy and paste' into new application areas. However, policy measures are likely to differ even more in the future, and evaluation will have to adapt. To highlight one key example, one major trend is the increasing importance of demand-driven innovation policy and diffusion-oriented measures. For these, evaluation practice is almost non-existent. This has a set of implications. Evaluation will have to tackle systematically and with methodological rigour a broader range of impacts – the focus on technological and economic impacts is increasingly too limited. Our understanding of how demand-side drivers and policies can interact with and influence supply-side developments also needs to improve radically before adequate evaluation approaches can be developed, and this understanding has to be shared by policy makers and evaluators alike.

A second example concerns the vastly increased emphasis the structural funds place on innovation, where there is a clear need for new innovation concepts in extremely challenging environments.[3] Without the development of intelligent and appropriate evaluation concepts and practices along the policy cycle, there is the danger that new application areas and innovation policy instruments might be supported by evaluation practices that are transferred without any consideration for contextual differences or – even worse – driven by ideological preconceptions. Hopefully, however, the lessons from INNO-Appraisal, the discourse we hope to support and the learning tool we provide can be of some assistance when designing and implementing improved and tailored evaluation approaches that will be needed in the future.

A final – and major – recommendation as to how the results of the study should be used relates to the repository that the study has designed and set up. This repository – in conjunction with the overall statistical data delivered in the study – is a comprehensive authoritative source that documents and codifies practices. The number of policy makers concerned with innovation policy and the number of analysts concerned with its assessment and improvement is constantly growing. Certainly there are guidebooks and manuals that describe evaluation concepts, methodologies and analytical techniques, and there is now an appreciable academic literature on evaluation, but the most numerous and useful sources of information – namely evaluation reports themselves – have to date been firmly embedded (some would say buried) in the relatively inaccessible 'grey literature'. INNO-Appraisal codifies much of the tacit knowledge that currently exists about evaluation practices and acts as a repository for this knowledge. It thus constitutes a source of learning for newcomers, a reference point for experienced practitioners and one way of helping to overcome problems associated with porous institutional memories. The INNO-Appraisal team strongly recommends that the EU Commission further invests in keeping the repository up-to-date, thus ensuring the survival of an institutionalised learning tool for evaluation and innovation policy in Europe. Moreover, INNO-Appraisal should be seen as a starting point for greater self-reflection by the evaluation community, with many more in-depth studies needed on evaluation practice and its contextualisation.

---

[3] First discussions between the INNO-Appraisal team and officials from DG Regio were held on February 4 2010 concerning the transfer and further development of concepts for structural fund evaluations in the area of innovation policy.

This page has been intentionally left blank.

PRO INNO EUROPE
INNO
APPRAISAL

Chapter I | Introduction

Innovation policy has become enormously diversified across Europe. The number of measures at national level has grown over the last few decades as more and more interventions seek to tackle different aspects of perceived market and system failures. The objectives of innovation policy have diversified, as have the designs of innovation policy interventions. Indeed innovation policy is in fact a mix of policies and is itself a more or less integral part of a broader policy portfolio at various levels. This development went hand in hand with a growing expectation as to the effectiveness of innovation policy to support capabilities and opportunities for innovation, growth and societal goals. What is less clear, however, is to what extent, and in what form, the ever growing aspiration of innovation policy is supported by appropriate analytical and formative means. The raising expectations of innovation policy do not seem to be met with an increasing sophistication in the use of evaluations to support policy and to better understand its effects.

For a range of basic questions we do not yet have the answer when it comes to evaluation practice in innovation policy in Europe:

How have those actors, who define the needs for an intervention, design programmes, and implement and adjust them over time, been using the tool box for evaluation across the whole policy cycle? What is the evaluation practice across Europe to underpin and justify the spending of public money on ever more increasing number of policy interventions? Are the sophisticated methods and analytical approaches developed within evaluation (e.g. Ruegg/Feller 2003, Fahrenkrog et al. 2002) utilised appropriately? Are evaluations built into innovation policy practice? Are tender and commissioning processes adequate, and how are evaluations discussed and used for the mobilising of stakeholders and learning across and beyond government? What can we say about the quality of evaluations, the discourse and consequences they trigger and their overall usefulness? Do we see a culture of policy-learning through the intelligent use of evaluation, and how is it influenced? What can we learn from the existing evaluation practice in innovation policy in order to further improve evaluation in efficient ways?

The project INNO-Appraisal has sought to contribute answers to these important questions. Its **first aim** has been to contribute to a **better understanding as to how evaluation is currently used in innovation policy in Europe**, and **how evaluation contributes to policy making**. INNO-Appraisal is the first systematic attempt to get an overview of evaluation practice in Europe. By doing so, it sought to achieve a **second, equally important aim**: i.e. to **make evaluation practice accessible** to the policy and evaluation community. In combination, the **third aim** then has been to contribute to a better informed and better networked evaluation discourse across Europe.

To achieve those aims, the project has for three years taken stock of and assessed evaluations in the area of innovation policy across Europe. It applied a novel and complex approach, combining qualitative in-depth analysis (case studies) and sophisticated quantitative analysis on the basis of a new form of data collection. The basis for the evaluation report collection has been the innovation policy database EU Trendchart in the period 2002 to 2007. The project designed and made use of a web-based template to allow a systematic characterisation of all selected evaluation reports. It then interacted with policy makers in order to verify and amend those characterisations. The template data was then used for the statistical analysis of the whole sample and of sub-samples for specific questions and case studies. To make those evaluation reports accessible, a repository of evaluation reports was created and put up on the INNO-Appraisal webpage. This repository allows interested parties to search for evaluation reports and actually to download them. It was also designed to allow a keyword search using the categorisation designed in the analytical template of each report. Thus, policy makers can now perform specific queries tailored to their specific need, be it – as examples – the application of certain methods, the coverage of certain topics or the evaluation of similar types of programmes.

The approach of the project, its various interim results and the repository itself have also been widely disseminated through the PRO—INNO® community and the wider policy and analyst community in innovation policy in Europe.

Thus, the major contributions of the project to the evaluation discourse and community in Europe are

(1) **An analysis of evaluation practice with some in-depth topical and country case studies (this report)**,
(2) the **repository** on the Inno-Appraisal webpage with all its various search and download functionalities and its function as a legacy of evaluation and a basis of a lasting stock-taking activity.

Together with the various interactions and presentations this project has already produced, and will continue to produce, this report (and earlier reports with interim results) as well as the repository should contribute to an **improved policy discourse**, of which all project participants sincerely hope it will continue to improve, and to which the legacy of this project can contribute. It needs to be stressed that the report is **not** another contribution to evaluation manuals (such as Fahrenkrog et al 2002; Ruegg/Feller 2003, OECD 1998, Miles/Cunningham 2006), it does not provide easy to apply lessons, but rather is analytical and provides a service to the Community in making evaluation in Europe more "tangible". The study used techniques of Meta-Evaluation,[4] but not to gather systematically how policy measures perform when combining the findings of different evaluations (as described in Georghiou 1999 and Edler et al. 2008), but to assess the overall design, implementation and functionality of evaluations **to learn about evaluation itself**, not about the impacts of the underlying policies (Implore 2009).

This report reflects the project as a whole. The **first part** of the report contains a detailed account of the overall methodology, including an explanation of the repository (chapter 2) and the overall analysis of the data (chapter 3). Chapter 3 summarises the general analysis of the evaluation repository. It provides an overall picture of the evaluation practice in Innovation Policy in Europe based on a descriptive analysis of our TrendChart database. It develops a first exploration of what determines evaluation design, implementation and effect and it explores if there are certain evaluation types (clusters) that could structure our future discussion on evaluation somewhat differently. In doing so it also lays the ground, in an exploratory fashion, for the in-depth case study to follow in subsequent chapters.

A **second part** of the study then contains four topical case studies, the selection of which has been done together with the steering committee:

---

[4] Edler et al. 2008 give some account of different approaches to Meta-Evaluation and Meta-Analysis.

Usefulness of evaluations (chapter 4), as the key final purpose of evaluation is to contribute to learning and improvement.

Measurement of impact (chapter 5), as the key of innovation policy is to make a difference and evaluation thus needs to be able to capture this.

Behavioural additionality (chapter 6), as a rather new, largely underexplored und often misunderstood concept of evaluation that has a high potential to improve policy making.

Evaluation in structural funds (chapter 7), as 20% of the evaluation reports are done in the context of the structural funds, whereby demands by the external (co-)sponsor as for evaluation clash with specific evaluation capabilities and institutions in the countries receiving money from structural fund. Moreover, as the structural funds now contain a large element of evaluation, the ability to evaluate, ex ante, interim and ex post, is becoming increasingly important, for the Commission and for the receiving countries.

A **third part** takes a country perspective, presenting four country case studies. The cases selected were **Austria (chapter 8)** as the most evaluation active country, **Germany (chapter 9)**, as a large country with systematic approaches but less variety and a less extensive culture of discussing evaluations broadly, the **UK (chapter 10)** that has a long history of evaluations, incubating in the past a set of key analytical approaches and finally, to cover Southern European approaches, a joint study on four countries: Portugal, Spain, Greece, Italy **(chapter 11)**.

The final part, **part four,** presents some reflections on the results of this study (**chapter 12**). It cannot and does not summarise all the findings. For that purpose, each of the individual chapter has its own executive summary and a separate executive summary to the study is provided at the front end of the study report. Rather, the final chapter provides some selected highlights from the study, puts the achievements of the project into the historical context of evaluation in innovation policy in Europe and concludes with an appeal to keep the stock taking and representation work started with the repository going – and with it to further improve and deepen the analysis of evaluation practice and its meaning for innovation policy in Europe more generally.

This project ran from January 2007 to January 2010. During all of this time a steering committee has overseen or followed the work, reflected on its various steps and come together to two key events. The advice given by the members of that committee has been extremely valuable. The project team would like to thank the policy makers Ulrike Blankenfeld (D), Mark Beatson (UK), Luisa Henriques (E) Nick Constantopoulos (EL), Rupert Pichler (A), Jari Romanainien (F), as well as our colleagues Aleardo Fulrani and Bart Kamp (INNOVA Europe), Phil Shapira (MBS, Manchester) and Anthony Arundel (MERIT)[5] very much for their advice and for their valuable time. Without this advice, the reassurance and constructive critique and the ideas for new avenues, the study team would have felt much less secure in pushing ahead with such a novel methodology.

We also would like to thank DG Entreprise for the opportunity to do such rewarding work. A specially heartfelt thank you goes to Alberto Licciardello, who has done a tremendous job as project officer to help to keep it all together and focused, to support in times of need, to coordinate with other services, to inspire the analysis and to disseminate our activities to colleagues inside and outside of the Commission. Finally, the project would like to thank all the participants at the INNO-Appraisal workshop in September 2009; the feedback gained at this event has been tremendously important and encouraging in the final stages of this project. Finally, our thanks goes to the numerous policy makers across Europe who have participated in this study by filling in templates about the evaluations they have commissioned, and by agreeing to be interviewed and otherwise giving feedback. Without this openness and responsiveness, the study would simply have been impossible.

---

[5] Because of changes in his working background Anthony Arundel had to drop out of the Steering Committee after a couple of months, unfortunatley.

References

Edler, J.; Ebersberger, B; Lo, V. (2008): Improving Policy Understanding by means of Secondary Evaluation; in R&D EVALUATION 17 (3), 175-186

Fahrenkrog, G., Polt, W., Rojo, J., Tubke, A., Zinöcker, K., and others (2002): RTD evaluation toolbox – assessing the socio-economic impact of RTD policies (EUR 20382 EN) Seville: IPTS. 2002. www.jrc.es/home/publications/publication.cfm?pub=1045)

Georghiou, L. (1999): Meta Evaluation. Evaluation of Evaluations, in Scientometrics 4(3), pp. 523-530.

Georghiou, L. (1995): Research evaluation in European National science and technology systems, in: Research Evaluation, 5(1), pp. 3-10.

ImpLore (2009). “Benchmarking Strategies and Methodologies of National, European and International R&D Programmes, to Assess and Increase their Impact on Innovation”, Report to Lot 2 of European Commission Tender ENTR/04/96. April 2009.

Miles, Ian, Paul Cunningham et al. (2005). “SMART Innovation: A Practical Guide to Evaluating Innovation Programmes”, A study for DG Enterprise and Industry. October 2005.

OECD (1998): Best Practice Guidelines for Evaluation. PUMA Policy Brief No. 5, May 1998 http://www.oecd.org/dataoecd/11/56/1902965.pdf. 1998.

Ruegg, R., and Feller, I. (2003): A Toolkit for Evaluating Public R&D Investment, Models, Methods, and Findings, from ATP’s First Decade. Gaithersburg, 2003.

Part I
Chapter 2

Methodology

This chapter summarises the methodological approach of the project. This approach was novel and complex, combining qualitative in-depth analysis (case studies) with a new form of data collection. Data search was organized centrally, but performed through all partners and a range of correspondents. It designed and made use of a web-based template to allow a systematic characterisation of all selected evaluation reports. It then interacted with policy makers in order to verify and amend those characterisations. The template data was then used for the statistical analysis of the whole sample and of sub-samples within case studies. A repository of evaluation reports was created and put up on the INNO-Appraisal webpage. Four case studies were conducted for selected topics – not individual evaluations – and four for selected countries or country groupings, the selection of those cases was discussed with the steering committee in two meetings. Their involvement as well as a workshop in September 2009 were integral parts of the methodological development, fine tuning and check.

Jakob Edler, Abdullah Gök and Paul Cunningham[6]

---

[6] The lead authors for this section of the report are as given above, with contributions from all other team members at different parts of its production, conclusions and interpretations have been approved and are shared by the group.

Table of Contents

References 6
Table of Contents 8
List of Tables 8
Table of Exhibits 8
1 Introduction 9
2 Build up and analysis of evaluation database 9
  2.1 Collection of quantitative data 10
    2.1.1 Template 10
    2.1.2 Coverage and correspondents 11
    2.1.3 The collection logic 13
    2.1.4 Quality assurance 16
  2.2 The innovation policy evaluation repository 16
  2.3 Statistical analysis 20
  2.4 In depth case study approach 20
  2.5 Steering Committee and Stakeholder Interaction 21

List of Tables

Table 1: Chronology of the Activities 10
Table 2: Structure of the Data Collection Template 11
Table 3: Consolidated Data Collection Statistics 12
Table 4: Sample Definitions 20
Table 5: Steering Committee – original composition 22

Table of Exhibits

Exhibit 1: Process Flow of the Data Collection 15
Exhibit 2: The repository on the webpage: The Landing Page 17
Exhibit 3: The repository on the webpage: The Country Page 18
Exhibit 4: The repository on the webpage: The Search Page 19

1 Introduction

The methodological approach within the project was novel and complex. The team invested huge time and effort to develop this approach. Its basic idea was to combine a new form of data collection, a template based survey that was used as a basis for collecting and displaying evaluation reports, characterising those reports and on that basis conducting quantitative and further in-depth quantitative analysis. The data search was organized centrally, but performed through all partners and a range of correspondents. It designed and made use of a web-based template to allow a systematic characterisation of all selected evaluation reports. It then interacted with policy makers in order to verify and amend those characterisations. The template data was then used for the statistical analysis of the whole sample and of sub-samples within case studies. A repository of evaluation reports was created and put up on the INNO-Appraisal webpage. Four case studies were conducted for selected topics and four for selected countries or country groupings, the selection of those cases was discussed with the steering committee in two meetings. Their involvement, as well as wider stakeholder involvement in a workshop in September 2009, were integral parts of the methodological development, fine tuning and check.

This section summarises and explains the methodological approach. It is substantive because the novelty and complexity of the approach needs a clear explanation. This approach, in combination with the web-based repository, is in itself an important outcome of the project. The web based repository and the template and process used to create it is one methodological and dissemination legacy that could be preserved and further developed. In order to do so, it is important to introduce the logic and the detailed application.

Part one explains the build up and analysis of the evaluation database, i.e. the collection of quantitative data through templates, introduces the repository of evaluations as it is on the webpage now, outlines the principles of the statistical analysis as well as the case study approach, and finally shortly explains the role of the steering committee and the wider interaction of INNO-Appraisal.

2 Build up and analysis of evaluation database

As summarised in Table 1, The INNO-Appraisal project started in March 2007 and ended in January 2010. A substantial part of the 3 years span of the project has been spent on the design and implementation of the data collection process. This part discusses the various issues in this process including the collection of quantitative data, the build-up of the publicly available innovation policy evaluations repository, the statistical and in-depth case study analysis conducted on the basis of the data collected.

It should be noted that as the project initially endeavoured to have a wider scope than evaluations by also including peer reviews, benchmarks etc., the earlier documents of the project including the data collection template uses the word “appraisal”. However, after a certain point, it was understood that the study should be limited to evaluation studies because of the limitations of the INNO-Policy Trendchart database on which the INNO-Appraisal project rely. Therefore, in all documents of the INNO-Appraisal project, including this report, the terms appraisal and evaluation are synonymous and interchangeable.

Table 1: Chronology of the Activities

Activity	**Timeframe**
Project Starts	March 2007
Preliminary Stock Taking	March-June 2007
Template Design	March – December 2007
First Manchester Workshop (internal)	July 2007
Pilot Applications of the Template	September-December 2007
Data Collection	January 2008 – October 2009
Data Analysis	
1st Round of Data Analysis	March – July 2008
Case Studies	September 2008 – January 2010
2nd Round of Data Analysis	April 2009 – July 2009
Vienna Workshop (internal)	July 2009
Final Round of Data Analysis	September 2009 – January 2010
Karlsruhe Workshop (internal)	January 2010
Workshops with steering committee and second level stakeholders	
First Steering Committee Meeting	July 2008
Brussels Workshop. Focus on Demand based innovation policy analysis (Lead Market Initiative)	January 2009
Brussels Workshop (with wider stakeholders and steering committee)	September 2009
Project Ends	January 2010

2.1 Collection of quantitative data

2.1.1 Template

The major empirical work in this study is the collection and structured description and analysis of evaluations. The first phase of the project was devoted to the preparation of data collection and analysis. It was agreed that a uniform template that could capture the variety of evaluation exercises in its major performance dimensions should be designed. It was also agreed that the template should be applied as strictly as possible and in a uniform manner across the board of evaluations. At the same time such a template would be the major means for the dissemination of data, Innovation Policy Evaluations Repository (IPER) which will be discussed later in this report.

The first main challenge for the design of a template was to keep a balance between being as analytical as possible and also to limit the template to a reasonable length. On the one hand, the project team felt that the template should be thoroughly analytical and comprehensive as it was the main data collection instrument of this project, and first of its kind as a comprehensive effort in the innovation policy domain. On the other hand, there are natural limits as to breadth of data collection through interaction with external partners. Therefore, it was decided that the template should be comprehensive but within reasonable limits to safeguard cooperation of policy makers and the necessary quality. The further challenges for the template were manifold. It needed to be comprehensible not only to the project team and the Commission, but to correspondents, policy makers, and the wider audience, from a broad diversity of national and linguistic backgrounds. It needed to capture different types of evaluations, and needed to be able to capture objective data (certain structural characteristics of evaluations) and subjective data (quality assessments) as well as information exclusively held by the policy makers (such as the consequences the evaluation had in the policy process).

The work on the template was challenging, very intensive and time consuming. Several iterations were done on the basis of a wide range of inputs, and a first internal pilot phase covering 8 evaluations was conducted with a preliminary working version of the template. On that basis the whole team met in Manchester in order to work on the template for two full days in July 2007. The template was finalised in December 2007. This process is summarised in Table 1.

The final structure of the template which is presented in Annex 2A includes the main elements listed in Table 2.

Table 2: Structure of the Data Collection Template

Part	Information collected/given
Cover page	Basic information on the project and instructions on how to fill in the template (for the policy makers)
Part A	Information about the policy maker responsible for the measure
Part B	Short information about the corresponding policy measure
Part C	Basis characteristics of the Evaluation: Who conducted it and tender procedure, Timing and purpose, Costs, Reason for the evaluation, Topics covered, Impacts analysed, Sponsor, Data analysis and collection methods used, Dissemination policies Main Audience
Part D	Quality of the Evaluation: Assessment of the methods and data (starting with terms of reference)
Part E	Recommendations of the Evaluation
Part F	Dissemination and Implementation of the Evaluation Results

2.1.2 Coverage and correspondents

The data collection process as summarised in Exhibit 1 started with the task of taking stock of the innovation policy measures that were reported evaluated in the INNO-Policy Trendchart database. From this database the team identified 293 evaluated innovation policy measures as summarised in Table 3. As the Trendchart database categorisation and search functionality allowed a query of “policy measures evaluated within last 5 years” and as some of the measures in the Trendchart database were last updated by late 2006 by the time the INNO-Appraisal team run the search, evaluations conducted between 2002 and 2007 (the year data collection started) were set as the scope of the project. 25 member states of the European Union, for whom the data was collected, were divided into 5 groups and each partner was assigned to one of them.

Consequently, the team derived the list of evaluations based on the Trendchart reported innovation policy measures evaluated between 2002 and 2007. As some policy measures were evaluated more than once and some evaluations covered more than one policy measure (i.e. portfolio evaluations), the team spent a considerable amount of time to derive the list evaluations from the Trendchart based list of evaluated innovation policy measures.

Table 3: Consolidated Data Collection Statistics

Country	EU25
Measures Reported as Evaluated in Trendchart	293
Evaluations as reported in Trendchart	352
Measures Added to Database	63
Evaluations Added to Database	124
Measures Dropped out of Database	169
Evaluations Dropped out of Database	214
Evaluations After Adding and Dropping	249
Evaluations for which there are Reports at Hand	242
Publishable Evaluations	173
Templates Filled in by Partners	154
Templates Filled in by Correspondents	88
Templates Sent to Policy Maker	216
Templates Returned from Policy Maker	146

We need to stress that the database **cannot** be interpreted as covering **all evaluations in innovation policy in the countries covered**. The data basis has been Trendchart. Even if the team corrected for some biases in the Trendchart database, a full coverage was not possible, nor was it essential for the main purpose of the study, i.e. to analyse how evaluations work (not to assess individual countries).- To illustrate the biases in the data, our preliminary list of measures as they appeared in the Trendchart database included the following two categories of biases, some of which could be remedied, others remained (at least partly):

- **Positive Bias:** The INNO-Appraisal team has found out that for some countries, the INNO-Policy Trendchart database listed more evaluated innovation policy measures as it actually was. For instance, for a country the INNO-Policy Trendchart database indicated 10 out of 30 measures were evaluated. However, consequently the country correspondent confirmed that none of them were really evaluated.
- **Negative Bias:** Similar to the above explained positive bias, it has been revealed that for some countries the INNO-Policy Trendchart database underreported the evaluated measures. The following examples illustrate this case:
    - For some countries, certain measures that were existent in the INNO-Policy Trendchart database and reported as not evaluated were revealed to have been evaluated. This was particularly the case for the measures that were evaluated recently, as the coverage of the INNO-Policy Trendchart database became quite outdated for a number of measures/countries.
    - For a number of countries the INNO-Appraisal team have identified evaluated innovation policy measures that were not covered in Trendchart. This is a fundamental bias as measures themselves were missed. The best example for these cases was Structural Funds where there were some types of evaluated innovation

policy measures within that were not covered in the INNO-Policy Trendchart database in full or part.

- Another problem was associated with the fact that the publicly available INNO-Policy Trendchart database was not retrospective. It only showed the situation as of the date it had been updated. However, the INNO-Appraisal project takes stock of evaluations that were conducted after 2002. Therefore, sometimes it proved extremely difficult to link the innovation policy measures in the INNO-Policy Trendchart database with evaluations. One particular example for this was that for a country, measures have been restructured (discontinued, changed name, merged, split, etc.) and the INNO-Appraisal team lost track of the evaluations of the old measures that were deleted from the database.

Further, the incoherency of the information contained in the INNO-Policy Trendchart database represented another challenge for the team. The information contained in the INNO-Policy Trendchart database with regard to evaluations, for instance, was incorrect in some cases. This was particularly evident for the website addresses of the evaluation reports presented in the database. Similarly, in some cases despite the database indicating there was no ex-ante evaluation (by having the expression “ex-ante evaluation: no”), the further information given by the database in the explanation section refuted this statement by mentioning specific ex-ante evaluations.

Finally, for a few countries, e.g. Finland and Sweden, the entries in Trendchart correspond to large agencies with a portfolio of different measures, but those measures often do not appear on the first level of Trendchart, and certainly not their evaluations. Evaluations in those countries are **under-represented** in our database.

Still, the team worked to minimise these biases which resulted in dropping of some measures and evaluations and also addition of some others. At the end of this process, summarised in Table 3, 249 evaluations were registered in the INNO-Appraisal database.

The INNO-Appraisal team gave a presentation to Trendchart Correspondents during their annual meeting in 2008 in Manchester and also prepared a report on the above discussed issues.

2.1.3 The collection logic

On the basis of the list of 249 evaluations as shown in Table 3, the INNO-Appraisal team collected the related evaluation documents. These documents included reports, report executive summaries ,and terms of references where available. This exercise was conducted by not only a search of publicly available sources but also constant communication with relevant Trendchart correspondents and respective policy-makers for some cases. Consequently, 242 evaluation documents were registered in the INNO-Appraisal internal repository.

For each of these 242 evaluations for which the team has the collected evaluation documents, a template discussed above was attempted to be filled in. It was decided that to ensure the level of harmony in the data collection, and also the quality-depth balance of the collected data, the template should first be partly pre-filled in by INNO-Appraisal partners. The characteristics evident from the report (parts A, B and C in the template) were identified by partners, and policy-makers were then consulted for the verification of this information (parts A, B and C in the template). Moreover, the policy makers were also asked to complete the information that was not evident from the report or other sources (parts D, E and F in the template) such as quality, usefulness and consequences. The broad category of policy-makers consisted of the following categories of individuals:

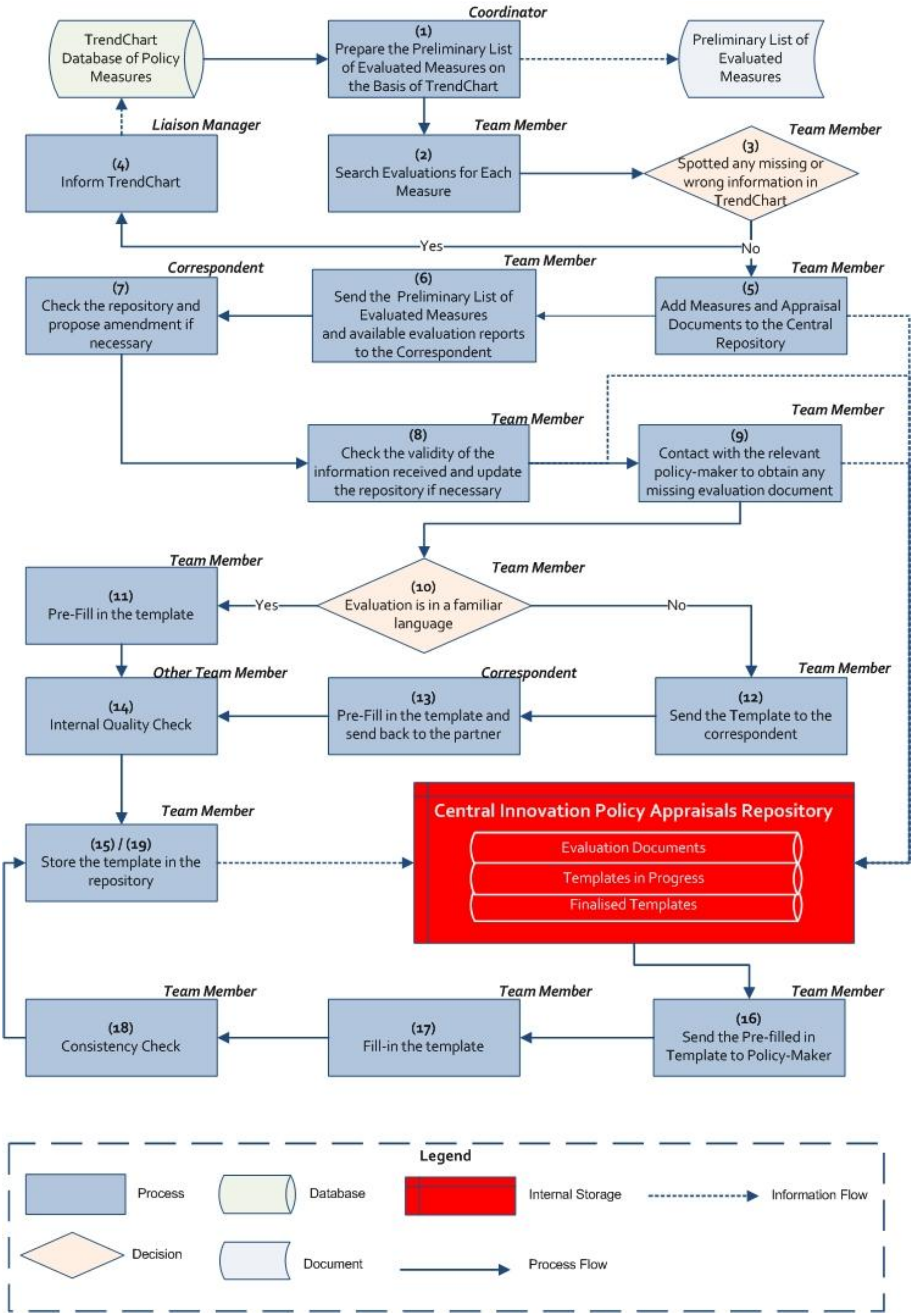
- Programme owners,
- Programme managers
- Commissioners of evaluation

As depicted in Exhibit 1, 242 evaluation reports were analysed and characterised by the above-explained INNO-Appraisal data collection template. For those countries where partners were familiar with the language of the report, partners pre-filled in templates on their own. Otherwise, respective Trendchart correspondents were utilised to pre-fill in templates. This process yielded a total of 242 pre-filled in templates.

Consequently, 242 pre-filled in templates were sent to respective policy makers for verification and completion. As it was not possible to locate some of the policy-makers due to of high degree of mobility of public officers, only 216 templates could be sent to respective policy-makers who had in depth experience and knowledge on respective evaluations. Finally, 146 templates were filled in by policy-makers and returned to the INNO-Appraisal team. This yields a 68% response rate.

The team had initially planned to utilise the Trendchart database for the analysis of the relationship between the evaluation and measure characteristics. However, after the evaluation data was collected, it was understood that the Trendchart database could not be used for measure characteristics. This was beacuse i) the Trendchart database had a major overhaul after the INNO-Appraisal project derived the list of measures from there which made the linking of evaluations to policy measures impossible and also ii) for those measures the INNO-Appraisal team could link evaluations to measures, the quality of information presented in the Trendchart database was considered as sub-optimal for a comprehensive analysis. Accordingly, the team decided to characterise the policy measure for which there was a template filled-in. This characterisation included the modality and the target group of measures.

Exhibit 1: Process Flow of the Data Collection


TrendChart Database of Policy Measures
Coordinator
(1) Prepare the Preliminary List of Evaluated Measures on the Basis of TrendChart
Preliminary List of Evaluated Measures
Liaison Manager
(4) Inform TrendChart
Team Member
(2) Search Evaluations for Each Measure
Team Member
(3) Spotted any missing or wrong information in TrendChart
Yes
No
Correspondent
(7) Check the repository and propose amendment if necessary
Team Member
(6) Send the Preliminary List of Evaluated Measures and available evaluation reports to the Correspondent
Team Member
(5) Add Measures and Appraisal Documents to the Central Repository
Team Member
(8) Check the validity of the information received and update the repository if necessary
Team Member
(9) Contact with the relevant policy-maker to obtain any missing evaluation document
Team Member
(11) Pre-Fill in the template
Yes
Team Member
(10) Evaluation is in a familiar language
No
Other Team Member
(14) Internal Quality Check
Correspondent
(13) Pre-Fill in the template and send back to the partner
Team Member
(12) Send the Template to the correspondent
Team Member
(15) / (19) Store the template in the repository
Central Innovation Policy Appraisals Repository
Evaluation Documents
Templates in Progress
Finalised Templates
Team Member
(18) Consistency Check
Team Member
(17) Fill-in the template
Team Member
(16) Send the Pre-filled in Template to Policy-Maker
Legend
Process
Database
Internal Storage
Information Flow
Decision
Document
Process Flow

2.1.4 Quality assurance

A Quality assurance process was in place from the very beginning. After the pilot implementation of the template, it was decided that uniform application of the template requires a shared understanding of the concepts by partners and also by correspondents and policy makers who contribute to the data collection exercise. To this end, an annotated template, in which the definitions of some of the concepts and further instruction included, was prepared.

Furthermore, at a later stage of the data collection exercise, the team implemented a test to determine the level of uniformity in data collection. A total of 10 randomly selected evaluations (2 from each partner's portfolio) were characterised by all other partners and the results were cross-checked. This test proved that the deviation between partners' understanding of the template was minimal and tolerable, especially as there was not one clear bias of partners in one or the other direction, but deviations were rather random and unsystematic.

2.2 The innovation policy evaluation repository

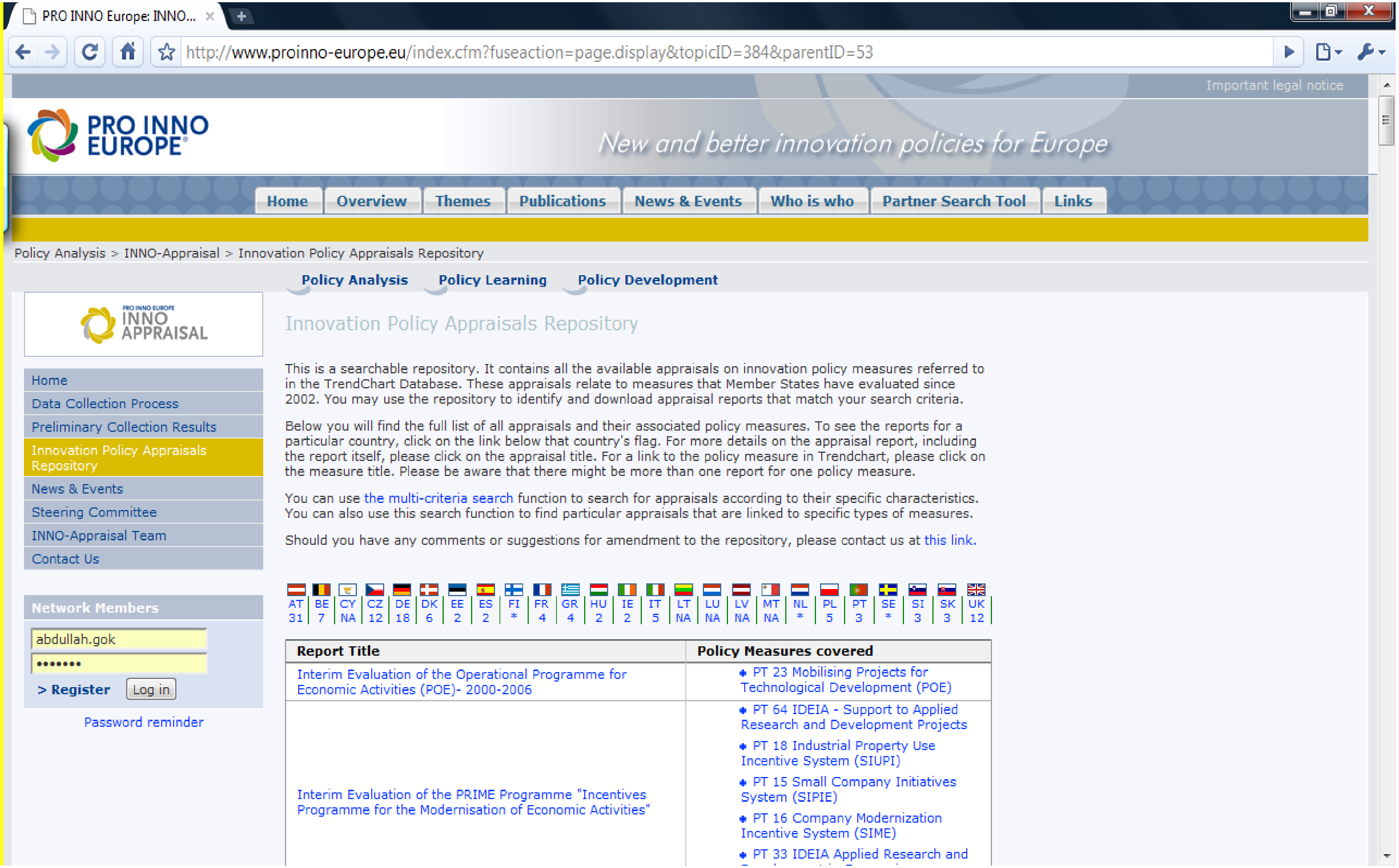
Evaluation reports collected by the INNO-Appraisal team forms a valuable source of information itself for policy-makers and policy analysts. After the collection and characterisation of these reports, the team published them in a public repository named Innovation Policy Evaluations Repository (IPER). These reports can be searched according to the following:

- Title of the evaluation or title of the related policy measure(s)
- Certain evaluation characteristics such as methods they employed and their timings etc.
- Characteristics of related policy measures.

The IPER is slightly more limited than the internal INNO-Appraisal internal repository as i) some of the evaluation reports were provided on the basis of confidentiality and also ii) evaluation characteristics that require subjective policy-maker opinion (i.e. quality and usefulness) should not be publicly published to ensure anonymity.

The IPER gathered considerable attention since its introduction on 11 May 2010. Within the first 2 months of its operation, until the end of June 2010, the INNO-Appraisal web site, including the IPER attracted 841 unique visitors. The IPER can be reached from www.proinno-europe.eu/appraisal. Exhibit 2, Exhibit 3, and Exhibit 4 below demonstrate the structure of the repository as of theend of January 2010.

Exhibit 2: The repository on the webpage: The Landing Page


PRO INNO Europe: INNO...
http://www.proinno-europe.eu/index.cfm?fuseaction=page.display&topicID=384&parentID=53
Important legal notice
PRO INNO EUROPE
New and better innovation policies for Europe
Home
Overview
Themes
Publications
News & Events
Who is who
Partner Search Tool
Links
Policy Analysis > INNO-Appraisal > Innovation Policy Appraisals Repository
Policy Analysis
Policy Learning
Policy Development
PRO INNO EUROPE INNO APPRAISAL
Home
Data Collection Process
Preliminary Collection Results
Innovation Policy Appraisals Repository
News & Events
Steering Committee
INNO-Appraisal Team
Contact Us
Network Members
> Register
Log in
Password reminder

abdullah.gok

•••••••

Innovation Policy Appraisals Repository

This is a searchable repository. It contains all the available appraisals on innovation policy measures referred to in the TrendChart Database. These appraisals relate to measures that Member States have evaluated since 2002. You may use the repository to identify and download appraisal reports that match your search criteria.

Below you will find the full list of all appraisals and their associated policy measures. To see the reports for a particular country, click on the link below that country's flag. For more details on the appraisal report, including the report itself, please click on the appraisal title. For a link to the policy measure in Trendchart, please click on the measure title. Please be aware that there might be more than one report for one policy measure.

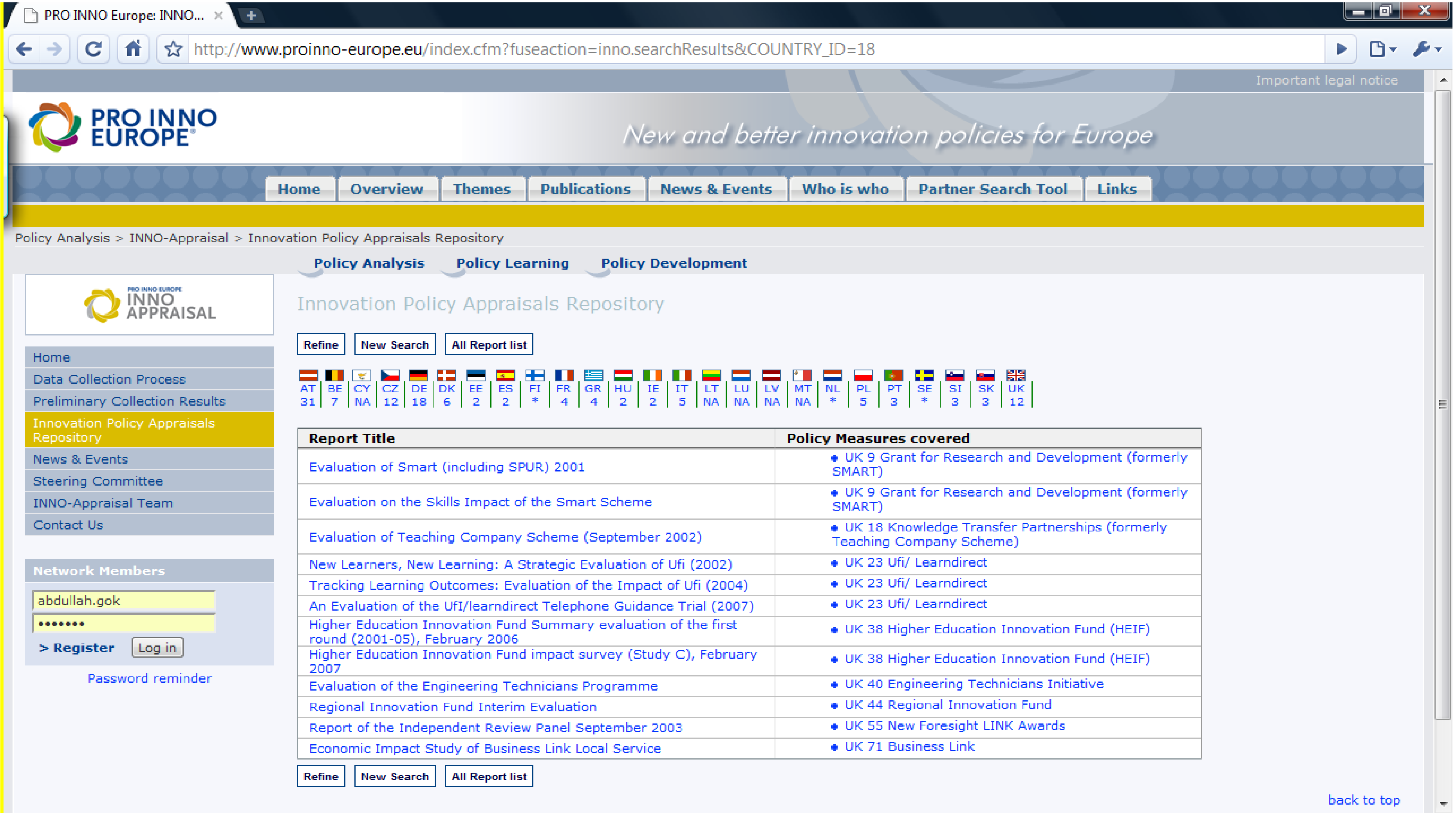
You can use the multi-criteria search function to search for appraisals according to their specific characteristics. You can also use this search function to find particular appraisals that are linked to specific types of measures.

Should you have any comments or suggestions for amendment to the repository, please contact us at this link.

AT	BE	CY	CZ	DE	DK	EE	ES	FI	FR	GR	HU	IE	IT	LT	LU	LV	MT	NL	PL	PT	SE	SI	SK	UK
31	7	NA	12	18	6	2	2	*	4	4	2	2	5	NA	NA	NA	NA	*	5	3	*	3	3	12

Report Title	Policy Measures covered
Interim Evaluation of the Operational Programme for Economic Activities (POE)- 2000-2006	◆ PT 23 Mobilising Projects for Technological Development (POE)
Interim Evaluation of the PRIME Programme "Incentives Programme for the Modernisation of Economic Activities"	◆ PT 64 IDEIA - Support to Applied Research and Development Projects ◆ PT 18 Industrial Property Use Incentive System (SIUPI) ◆ PT 15 Small Company Initiatives System (SIPIE) ◆ PT 16 Company Modernization Incentive System (SIME) ◆ PT 33 IDEIA Applied Research and

Exhibit 3: The repository on the webpage: The Country Page


PRO INNO Europe: INNO...
PRO INNO EUROPE
New and better innovation policies for Europe
Important legal notice
Home
Overview
Themes
Publications
News & Events
Who is who
Partner Search Tool
Links
Policy Analysis
Policy Learning
Policy Development
INNO APPRAISAL
Home
Data Collection Process
Preliminary Collection Results
Innovation Policy Appraisals Repository
News & Events
Steering Committee
INNO-Appraisal Team
Contact Us
Network Members
> Register
Log in
Password reminder
Refine
New Search
All Report list
back to top

http://www.proinno-europe.eu/index.cfm?fuseaction=inno.searchResults&COUNTRY_ID=18

Policy Analysis > INNO-Appraisal > Innovation Policy Appraisals Repository

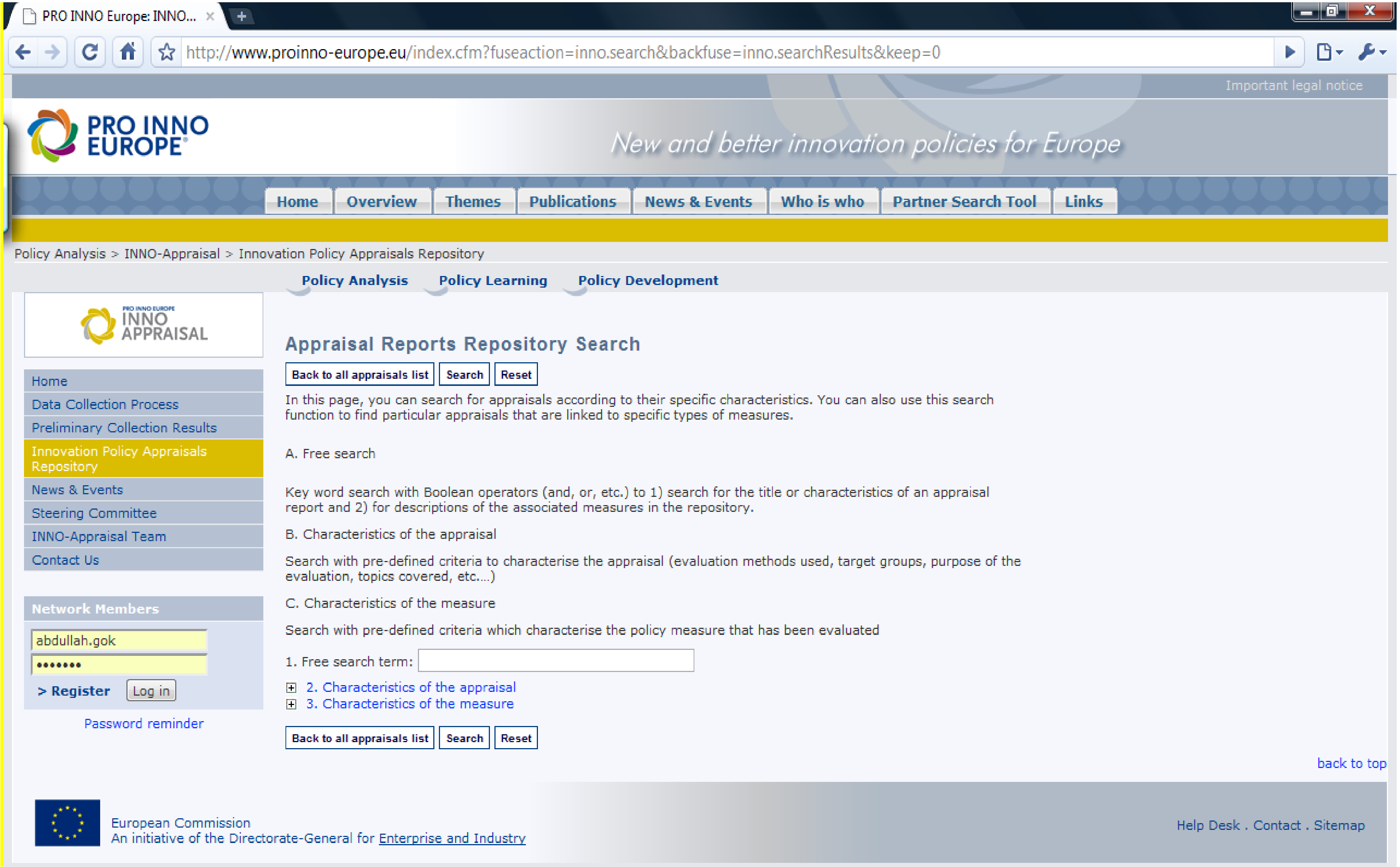
Innovation Policy Appraisals Repository

AT	BE	CY	CZ	DE	DK	EE	ES	FI	FR	GR	HU	IE	IT	LT	LU	LV	MT	NL	PL	PT	SE	SI	SK	UK
31	7	NA	12	18	6	2	2	*	4	4	2	2	5	NA	NA	NA	NA	*	5	3	*	3	3	12

Report Title	Policy Measures covered
Evaluation of Smart (including SPUR) 2001	UK 9 Grant for Research and Development (formerly SMART)
Evaluation on the Skills Impact of the Smart Scheme	UK 9 Grant for Research and Development (formerly SMART)
Evaluation of Teaching Company Scheme (September 2002)	UK 18 Knowledge Transfer Partnerships (formerly Teaching Company Scheme)
New Learners, New Learning: A Strategic Evaluation of Ufi (2002)	UK 23 Ufi/ Learndirect
Tracking Learning Outcomes: Evaluation of the Impact of Ufi (2004)	UK 23 Ufi/ Learndirect
An Evaluation of the UfI/learndirect Telephone Guidance Trial (2007)	UK 23 Ufi/ Learndirect
Higher Education Innovation Fund Summary evaluation of the first round (2001-05), February 2006	UK 38 Higher Education Innovation Fund (HEIF)
Higher Education Innovation Fund impact survey (Study C), February 2007	UK 38 Higher Education Innovation Fund (HEIF)
Evaluation of the Engineering Technicians Programme	UK 40 Engineering Technicians Initiative
Regional Innovation Fund Interim Evaluation	UK 44 Regional Innovation Fund
Report of the Independent Review Panel September 2003	UK 55 New Foresight LINK Awards
Economic Impact Study of Business Link Local Service	UK 71 Business Link

abdullah.gok
•••••••

Exhibit 4: The repository on the webpage: The Search Page


PRO INNO Europe: INNO...
http://www.proinno-europe.eu/index.cfm?fuseaction=inno.search&backfuse=inno.searchResults&keep=0
Important legal notice
PRO INNO EUROPE
New and better innovation policies for Europe
Home
Overview
Themes
Publications
News & Events
Who is who
Partner Search Tool
Links
Policy Analysis > INNO-Appraisal > Innovation Policy Appraisals Repository
Policy Analysis
Policy Learning
Policy Development
PRO INNO EUROPE INNO APPRAISAL
Home
Data Collection Process
Preliminary Collection Results
Innovation Policy Appraisals Repository
News & Events
Steering Committee
INNO-Appraisal Team
Contact Us
Network Members
> Register
Log in
Password reminder
Back to all appraisals list
Search
Reset
back to top
European Commission
An initiative of the Directorate-General for Enterprise and Industry
Help Desk . Contact . Sitemap

Network Members: abdullah.gok / •••••••

Appraisal Reports Repository Search

In this page, you can search for appraisals according to their specific characteristics. You can also use this search function to find particular appraisals that are linked to specific types of measures.

A. Free search

Key word search with Boolean operators (and, or, etc.) to 1) search for the title or characteristics of an appraisal report and 2) for descriptions of the associated measures in the repository.

B. Characteristics of the appraisal

Search with pre-defined criteria to characterise the appraisal (evaluation methods used, target groups, purpose of the evaluation, topics covered, etc.…)

C. Characteristics of the measure

Search with pre-defined criteria which characterise the policy measure that has been evaluated

1. Free search term:

⊞ 2. Characteristics of the appraisal

⊞ 3. Characteristics of the measure

2.3 Statistical analysis

As a consequence of the data collection process, the INNO-Appraisal analysis utilizes two different samples, each tailored for a specific purpose (Table 4). This procedure is necessary, because first, not all questions can be referred to the total sample, and second, the raw dataset exhibits certain biases, which need to be corrected. For example quality, usefulness, as well as impacts of evaluations can only be analysed for returned templates, since these are the only ones that were filled in/ reviewed by policymakers. Finally, for particular countries the dataset includes several evaluations which are part of the same portfolio/ structural fund evaluations. Naturally, these contain identical information (at least on the characteristics of the evaluation, because these relate to the total portfolio/ structural fund evaluation). Including this raw data would cause biased inflation of data. Consequently, it was necessary to correct for this type of evaluation in the samples. On the other hand some of these sub-evaluations were commissioned (and consequently reviewed) by different policy makers which made it necessary to include them in the analysis of the quality aspects etc.

Table 4: Sample Definitions

Sample	Definition	Potential use	n	remark
A	Includes all received questionnaires as well as the prefilled templates if no template was received;	Characteristics	171	GR, PT: since many individual evaluations are part of larger structural fund evaluations, Atlantis provided one ‘umbrella’ observation for each SF in order to avoid a country bias
B	Only received templates	Quality and Recommendations	132	GR, PT: individual templates of SF evaluations were used, because these were reviewed by different policy makers

By using this data, firstly, an analysis of the basic features of the INNO-Appraisal data was conducted. This part utilised mostly descriptive statistics. Consequently, the team, in close collaboration with the Steering Committee, has devised a number of hypotheses. The test of them constituted the second stage of the quantitative analysis. Chapter 3 of this report is an account of the statistical analysis of both samples, and other parts of the study also draw on this database.

2.4 In depth case study approach

A further pillar of the methodology were case studies. The idea of case studies was to have some in-depth analysis on specific topics and on specific countries in order to understand better how evaluations work in certain context conditions (country cases) and how certain aspects of evaluation can be better understood when analysing the data specifically, and in light of existing literature on those topics

A first selection of country case studies was undertaken by the team on the basis of variety for certain variables and sufficient data availability. The cases selected were Austria, as the most evaluation active country, Germany, as a large country with systematic approaches but less variety and a less extensive culture of discussing evaluations broadly, the UK that has a long history of evaluations, incubating in the past a set of key analytical approaches, and finally, to cover Southern European approaches, a joint study on four countries: Portugal, Spain, Greece, Italy. Initial versions of those studies were discussed with the steering committee in the first meeting in 2007.

As for the topics case studies, the group again produced a short list of potential topics that were understood as being interesting and of value to the project. This list was then electronically discussed with the steering committee and the Commission. The final selection is the result of this feedback process.

The four cases selected were:

Usefulness of evaluations, as the key final purpose of evaluation is to contribute to learning and improvement

How to measure impact?, as the key of innovation policy is to make a difference and evaluation thus needs to be able to capture this

Behavioural additionality, as a rather new, largely underexplored und often misunderstood concept of evaluation that has a high potential to improve policy making

Evaluation in structural funds, as 20% of the evaluation reports are done in the context of the structural funds, whereby demands by the external (co-)sponsor for evaluation clash with specific evaluation capabilities and institutions in the countries receiving money from structural fund. Moreover, as the structural funds now contain a large element of evaluation, the ability to evaluate, ex ante, interim and ex post, is becoming increasingly important, for the Commission and for the receiving countries.

The case studies were conducted through literature reviews, web-based document research, analysis of the INNO-Appraisal database (and the related templates) and a set of in-depth interviews with evaluators and policy makers. In the September 2009 project workshop, the first versions of the case studies were discussed with the participants and feedback integrated for the final versions contained in this report.

2.5 Steering Committee and Stakeholder Interaction

As just outlined above, INNO-Appraisal had a steering committee for support and feedback. The steering committee met in person twice, in critical phases of the project (see above, Table 1), but were also in communication through electronic exchanges at all important steps of the project. Of course, the usual disclaimer remains, all decisions taken in the project are soly the responsibility of the project team.

The steering committee was composed as shown in Table 5. This took into account a geographical diversity and different policy functions as well as academic background. Due yo a set of personal developments, Anthony Arundel (changing jobs) and Philipp Shapira (joining the MIoIR) only accompanied the project in the first half of its existence, although Philip Shapira remained connected to the project through his new affiliation. Additionally, Arleado Furlani was replaced by Bart Kamp.

Table 5: Steering Committee – original composition

Group	Name of the Steering Committee Member	Affiliation
Policy Makers	Ulrike Blankenfeld	BMWi, Germany
	Mark Beatson	Department of Trade and Industry, UK
	Luisa Henriques	Fundação para a Ciência e Tecnologia. Portugal
	Nick Constantopoulos	General Secretariat for Research and Technology, Greece
	Rupert Pichler	BMVIT, Austria
	Jari Romanainen	TEKES, Finland
Analysts	Phil Shapira	Manchester Institute of Innovation Research (formerly Georgia Institute of Technology
	Anthony Arundel	United Nations University – MERIT
	Aleardo Furlani	INNOVA Europe

As for interaction with stakeholders beyond the steering group, a set of methods and channels were applied. The following list simply names the major venues, it does not give all interaction details

- One open workshop in September 2009, with an impressive turnout
- Presentation within the PRO INNO Europe Community (Trendchart Workshop)
- The implementation of the Website, as traditional means of communication, but also as a host of the repository (see above), fully integrated within the PRO INNO Europe website
- Four newsletters summarising major events and highlights, widely published.
- A whole set of bilateral exchanges, such as integration into OMC Net, discussion with other Commission units (DG Research, DG Regio), and discussion at national level ministries (UK, Austria, Germany etc.)

Part I
Chapter 3

The Use of Evaluation in Innovation Policy in Europe – A Statistical Analysis

This chapter is the main, general analysis of the evaluation repository. It provides an overall picture of the evaluation practice in Innovation Policy in Europe based on a descriptive analysis of our Trendchart database. It develops a first exploration of what determines evaluation design, implementation and effect and it explores if there are certain evaluation types (clusters) that could structure our future discussion on evaluation somewhat differently. In doing so it also lays the ground, in an exploratory fashion, for the in-depth case study to follow in subsequent chapters.

Jakob Edler, Abdullah Gök, Martin Berger, Michael Dinges
in collaboration with the rest of the project team

Table of Contents

Table of Contents 24

List of Tables 25

Table of Exhibits 26

Executive Summary 27

1 Introduction 31

2 Use and characteristics of evaluation in innovation policy in Europe 31

  2.1 Basic characteristics of the dataset 31

  2.2 Evaluation characteristics 36

3 Exploring and Understanding Evaluations – some cross-cutting analysis 50

  3.1 The meaning of policy measures 51

  3.2 Characterising evaluation practice – determinants of evaluation design 56

    3.2.1 Introduction 56

    3.2.2 Timing of evaluations 56

    3.2.3 Purpose of evaluations: formative vs summative 56

    3.2.4 The meaning of co-sponsors 60

    3.2.5 The link of topics and methods 62

    3.2.6 The link of topics with methods and data collection approaches 63

  3.3 Consequences and quality of evaluations 70

    3.3.1 Determinants and consequences of evaluation quality 70

    3.3.2 Determinants of evaluation consequences 78

  3.4 Cluster of evaluations – an exploration 83

References 85

List of Tables

Table 6: Sample Definitions 32
Table 7: Correlation between modality of policy measure and quality and usefulness........ 54
Table 8: Timing of the evaluation and type of evaluation 57
Table 9: Topics covered by type of evaluation (summative vs. formative) 58
Table 10: Data analysis methods used by type of evaluation (summative vs. formative) 59
Table 11: Data collection method and sources by type of evaluation (summative vs. formative) 60
Table 12: Topics covered and externals sponsorship (multiple responses) 61
Table 13: Impacts covered (aggregated) and externals sponsorship of the evaluated measure 62
Table 14. Share of evaluations covering a specific combination of topics and impacts........ 63
Table 15. Share of evaluations covering a specific combination of topics and methods (pairwise) 66
Table 16. Share of evaluations covering a specific combination of topics and data collection approaches (pairwise) 68
Table 17: Correlation Matrix quality indicators (pairwise) 71
Table 18: Evaluation condition of an external/ international (co)sponsor and evaluation quality...... 72
Table 19: Sponsor of the evaluation and evaluation quality 73
Table 20: Tender procedure and evaluation quality........ 73
Table 21: Median of quality aspects by timing of the evaluation........ 75
Table 22: Main intended audience and evaluation quality 76
Table 23: Correlation coefficients between quality and breadth of discussion (Spearman, pairwise) 78
If we look at the composite quality indicator, the only statistical significant difference is that high quality evaluations lead more often to the expansion/ prolongation of a measure (see Table 24). ... 80
Table 25: Correlation coefficients between individual quality indicators and consequences of the evaluation (Spearman, pairwise) 82
Table 26: Correlation coefficients between discussion indicators and consequences of the evaluation (Spearman, pairwise) 83
Table 27: Comparison of two clusters of evaluation 84

Table of Exhibits

Exhibit 5: The Type of Sources of Information for the Collected Data 32
Exhibit 6: Types of Policy Measures Associated with Evaluations 33
Exhibit 7: Target Groups of Policy Measures Associated with Evaluations 34
Exhibit 8: The Share of Structural Fund Evaluations 34
Exhibit 9: The Share of Portfolio Evaluations 35
Exhibit 10: Composition of Structural Fund and Portfolio Evaluations 35
Exhibit 11: Country Distribution in the INNO-Appraisal dataset 36
Exhibit 12: Type of Evaluator 36
Exhibit 13: Tender Procedures of Evaluations 37
Exhibit 14: Timings of Evaluations 37
Exhibit 15: Purpose of Evaluations 38
Exhibit 16: Evaluations Sponsored Externally 38
Exhibit 17: Evaluation Budget 38
Exhibit 18: Evaluation Planning 39
Exhibit 19: Evaluation Topics 40
Exhibit 20: Clustering of Topics 41
Exhibit 21: Impacts Looked at in Evaluations 42
Exhibit 22: Evaluation Sponsors 42
Exhibit 23: Data Analysis Methods and Main Evaluation Designs/Approaches Employed in Evaluations 43
Exhibit 24: Data Collection Methods and Data Sources Employed in Evaluations 44
Exhibit 25: Availability of Evaluations Terms of References as Part of Report 44
Exhibit 26: Availability of Evaluations Terms of References in Other Sources 45
Exhibit 27: Clearly Stated Objectives in Evaluation Terms of References 45
Exhibit 28: Specification of Methodologies and Approaches in Terms of References 45
Exhibit 29: Main Intended Audiences of Evaluations 46
Exhibit 30: Discussions of Evaluations 46
Exhibit 31: Perceived Quality of Evaluations 47
Exhibit 32: Recommendations in Evaluations 48
Exhibit 33: Perceived Usefulness of Recommendations of Evaluations 49
Exhibit 34: Consequences of Evaluations 50
Exhibit 35: Distribution across quality categories 70
Exhibit 36: Type of tender and evaluation quality criteria 74
Exhibit 37: Distribution of width of discussion and evaluation quality 77
Exhibit 38: Width of discussion by type of evaluation in % (n in brackets) 79
Exhibit 39: Consequences by type of evaluation (in %; category other omitted) 80

Executive Summary

The INNO-Appraisal repository that has been built up over three years allows the statistical analysis of evaluation practice in Europe. This chapter delivers a characterisation and analysis of evaluation in innovation policy across Europe in a very general sense, to deliver a picture of practice, quality and consequences of evaluations in innovation policy.

The characteristics of evaluations in the INNO-Appraisal database

The INNO-Appraisal database contains evaluation of a whole range of different policy measures that are covered in the Trendchart Database between 2002 and 2007. Reflecting the innovation practice across Europe, its majority of evaluations are concerned with direct financial support for innovation activities, and two thirds of the underlying sample measures in the database are geared towards involving Universities and public research institutes. The database of evaluations covers all European countries, with an interesting bias towards Austria which has an exceptionally high number evaluation innovation policy measures reported in Trendchart and an extensive evaluation activity. The repository, as basis for the analysis, cannot claim to cover all innovation policy evaluation in all countries to the same degree, as countries represent their activities differently in the Trendchart database, the basis of the analysis. For example, Finland and Sweden are underreported in the database as many of their evaluations are done within the programme portfolio of large agencies so that many individual measures are not flagged out in the Trendchart database. To get an understanding of the meaning of country contexts, however, later sections deliver in-depth country cases of Austria, Germany, UK and Mediterranean countries. The repository also covers extensively evaluation of structural fund measures, as slightly more than 20% of all evaluations are performed in the context of structural funds. The **number of evaluation reports** altogether is **242**, of which **216** could be meaningfully analysed by the project team (and thus used for the statistical analysis presented), and **146** were amended and **verified by policy makers** (used for specific, judgemental and policy related parts of the statistical analysis). The number of publishable evaluation reports in the repository of the project is 173.[7]

Commissioning and design: Evaluation is found to be an integral part of innovation policy, as roughly **50% of the measures** that are evaluated have a **pre-determined budget** for evaluation and **two thirds** are **foreseen and planned** in the measure design. More than 90% of evaluations are sponsored by the programme owners themselves, only a minority are jointly sponsored with other bodies or entirely external (10%). **Almost half** of the evaluations follow an **open tender procedure**, one fifth are done through closed tender, one fifth are performed by external evaluators without a tender and 15% are done internally. For those evaluations having **a tender**, a **large majority clearly specified the objectives**, at the same time, two thirds of the tender documents left the choice of methods to the evaluators in.

Timing: More than 40 % of the database are interim evaluations. This bias against ex post (30%), however, stems partly from the selection method focusing on live Trendchart policies within a certain period of time. The database contains both formative (33%) and summative (21) evaluations, while **the majority combines both summative and formative aspects**.

---

[7] For a series of methodological and database specific reasons one cannot give a statistical data as for the share of policy measures that are evaluated within Trendchart in the period covered.

Topics: The topics covered in evaluations are broad, obviously. In very general terms, **effectiveness** and **consistency** appear to be **slightly more important** than programme efficiency issues, while the in-depth look at **project efficiency** is **much less common** (below 50%). We also find a certain **clustering of topics** that are covered. Two thirds of all evaluations cover at least one form of additionality (input, output and behavioural), and many of those evaluations tend to include the project level in order to understand those additionalities. Gender (24%) and minority (7%) are least common. In terms of impact, technological and economical are most important, and environmental impacts (still) least important (28%).

Methodology and data sources: In terms of methodology, we find a whole range of methods applied, however, some general strong trends are obvious. **Descriptive statistics** are the **most common** approach, applied by more than three quarters of all evaluations, while case studies – to understand contexts and developments over time – are performed only by 41%. **More sophisticated,** quantitative approaches are used **even more selectively**, e.g. 23% perform econometric analysis, 17% network analysis. Interestingly, 80% claim to use monitoring data and 70% to use existing surveys and databases as a basis for the analysis. However, it appears that this kind of data is insufficient to be used for specific evaluation questions such as networking or behavioural additionality. The most important *pro-active* **data collection is done through interviews and participant surveys**. **Technometric analysis** in innovation policy plays **no significant role at all** (2%), it appears that for the analysis of technological substance in projects peers are used (20%).

***Quality*:** As for overall quality of evaluations, the database shows very mixed results along nine different quality aspects. For a general picture a simple binary quality index has been constructed, all evaluations that score more than 3 on a Likert scale (1 being very low, 5 being very high) in *each* of *four* selected quality variables are defined as being of high quality. 61% of the evaluations show an overall positive quality index. This means that **almost 40% of the evaluations have serious quality problems in at least one key quality dimensions**. This finding is confirmed through an auto-correlation analysis: Many evaluations are either good in a whole set of quality criteria or perform rather badly across the board.

The policy use of evaluations: While almost all evaluations are targeted towards policy makers and programme management, **only 50**% of the evaluations are **targeted towards the users** of the programme and **less than one third to the general public**. Evaluations are obviously **not extensively used to mobilise** the community, policy makers themselves rate the breadth and depth of actual discussion about evaluation results only moderate.

Most evaluations, as to be expected, do contain recommendations for policy and programme management, **only a minority of evaluations is purely analytical**. The **usefulness of the recommendations** for various aspects of policy learning and improvement that were tested is **moderate** and appears to have room for improvement. In principle, **evaluations are not linked with major, radical consequences**, those appear to be the result of more general policy considerations. However, they are **important for minor re-**design of measures or their **prolongation and extension**. In 17% of all cases they are also used to improve other or future policy measures.

Determinants of evaluation practice, quality and consequences

There is a certain degree of convergence of evaluation practice across different policy measure, we find **surprisingly little variation between different policy measures** as regards a **whole range of evaluation characteristics**, such as tender procedures, internal vs. external evaluators, coverage of topics and impacts and even use of some of the data collection approaches and methods and even targeted audiences. It shows that other factors, such as organisational and country specific traditions, topics to be covered and general practices dominate the design and implementation of evaluations to a large extent, not so much the evaluation object – the policy measure – itself.

However, the **type of measure makes *some* difference** as for evaluation design and implementation. Certain specific types of programmes show a specific application of tailored methods and data collection approaches (e.g. network analysis and case study approaches for networking and cluster programmes). We also find variation in the **use and dissemination of evaluation** between policy measures, with - for example – **complex networking programmes targeting beneficiaries much more often** as those measures are complex and need explanation and formation. Furthermore, evaluations for direct financial support measures and for cluster, technology transfer and networking measures are more likely to be perceived as being of good quality, while **evaluations for softer measures** such as **management support measures** or **diffusion measures** are of **lower quality**. In addition, there seems to be a **poorly developed evaluation practice for diffusion measures**, which – in addition – do not take societal and environmental impacts into account as broadly as to be expected, and that are perceived to e of less usefulness to policy makers.

Evaluations are often **influenced by external sponsors** of the policy measures. While they do not impose methods they introduce a **bias towards social and environmental impacts and gender** and minority issues. The external sponsors, it seems, are one major reason behind a **certain grouping** of evaluations around topics we observe, some being more concerned with quantitative, hard economic and technological outputs, and others interested in social, environmental impacts etc.

Evaluators in innovation policy appear to apply a **form follows function** approach, they tailor their approaches according to the need for topics and impacts to be covered. For example evaluations interested in strategy development and policy issues more general also look at consistency and use vastly interviews and other qualitative methods. Evaluations more concerned with **effectiveness** rely on (often simple) **statistical analysis and data**, and the **use of peers**, although limited, is **strongly linked to quality of output**. Those evaluations more concerned with **efficiency and project level issues**, in turn, tend to look for different kinds of **additionality** and rely on surveys, interviews and, less broadly, though, on case studies. Further, formative and ex ante evaluations tend to analyse consistency issues more broadly than other evaluations (i.e. to assess and re-adjust the overall match), and they do so be using slightly more qualitative methods.

A deeper look into the **determinants of quality assessments** reveals that policy makers see **room for improvement** as regards the **coverage of the broader context**, the **application of advanced quantitative** and some qualitative methods and the **documentation of information sources**. In contrast, evaluations covering **technological and scientific impact** and those using survey methods and peer review are perceived of being of higher quality. Summative evaluations appear to be perceived as being of higher quality than formative evaluations, and indeed they are more widely discussed within government than formative ones. Formative evaluations, it seems, are a tool for improvement for the programme owners and beneficiaries, while the messages of summative evaluations are used for wider discourse and justification.

Interestingly, **quality** does not differ between evaluations that are done by external evaluators and those performed internally. Equally, evaluations are **not** perceived to be of **higher quality** if they are **in-built in policy measures from the start** and **have a dedicated budget** within the policy measure. However, one important finding is that quality is lower for evaluations that are commissioned by external sponsors or policy bodies. In contrast, open tenders yield evaluations with better quality.

Quality, finally, **makes a difference** when it comes to the **dissemination and exploitation** of evaluations. The better an evaluation, the more likely it is discussed within and outside government. In addition, evaluations that ex ante are targeted to the wider public and policy analysts (and not only to the programme management) are also of higher quality.

The analysis also revealed that evaluations have **a limited set of consequences**, **radical consequences** (termination of programmes) are **very rarely a result of an evaluation**, but rather they appear to be consequence of principle policy decisions. In contrast **evaluations lead to minor re-design of measures or learning for other measures** and, most often, to **prolongation and extension**. The latter is highly correlated with simple methods, it thus appears that **clarity and simplicity** in the data and methods is part of a **confirmation and incremental approval exercise**. In addition, those evaluations which are intensively discussed within and outside government are those that are more likely to lead to consequences. Finally, quality also is important for evaluation consequences, evaluations of higher quality more often tend to lead to consequences (especially prolongation). The quality aspects most strongly linked to the likelihood for consequences out of the evaluations are the extent to which evaluation methods satisfy the Terms of Reference and the purpose of the evaluation.

In a final analytical step the general statistical analysis explored **clusters of evaluations**. Two clusters emerge. One cluster of evaluations is more populated by ex ante evaluations and is concerned with programme efficiency issues and, by nature, more often based on qualitative methods. The second cluster appears to be more ex post and interim, being broader in its coverage and more concerned with different forms of outcome/impact, thereby mobilising more quantitative approaches and oriented towards the policy community rather than the beneficiaries. This cluster of evaluations is more often used for decision about prolongation or re-design of measures.

1 Introduction

Based on the data obtained through the methodology described in the previous section, this section delivers a statistical analysis to understand evaluation practice in innovation policy in Europe. The first section presents the results of a descriptive analysis of our database, thereby providing a picture of the relative frequency with which evaluations are used for innovation policy and of the distribution of specific evaluation practices and characteristics of the evaluation. The main characteristics of evaluations to be presented concern

- Tender procedures,
- Conductor of the evaluation,
- The planning of evaluations within policy measures (budget, ex ante planning)
- Timing and purpose
- Topics covered
- Impacts covered
- Sponsorship of the policy measure and the evaluation
- Publication, language and availability of the report
- Data analysis methods used
- Data collection methods
- Intended audiences
- Quality aspects (assessment of quality)
- Concussions and recommendations

In a further step we test for the relationships between variables in order to understands the design of evaluations better. To do so, we first relate the various characteristics of evaluations to a typology of policy measures and target groups. The idea behind this analysis is that different types of policy measures (for different kinds of target groups) might use evaluations differently. On that basis, section 3 subsequently analysis how different characteristics of evaluations relate to each other in order to better understand the nature of evaluations, the co-existence of different characteristics pointing towards different types of evaluations. This section will only shortly report on those issues that have been selected for in-depth case studies, such as behavioural additionality, impact, structural fund evaluations and usefulness of evaluation. Finally, section 3 will end with an exploratory cluster analysis to see if the high number of evaluations shows some pattern, some similar types of evaluation.

2 Use and characteristics of evaluation in innovation policy in Europe

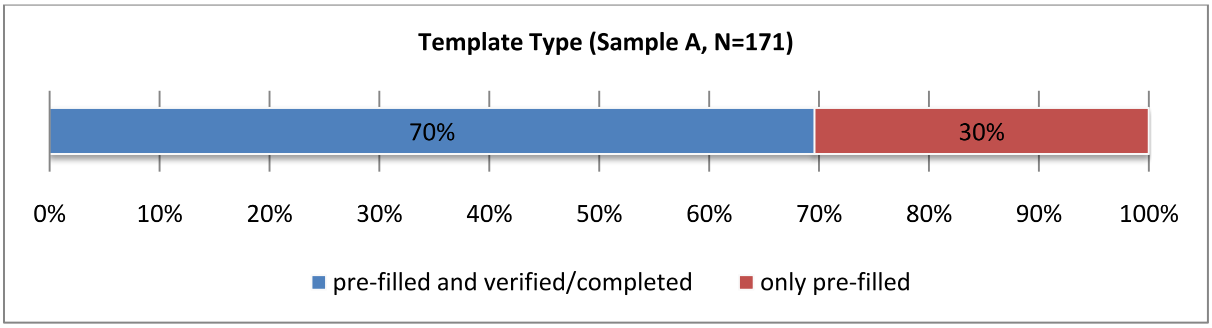
2.1 Basic characteristics of the dataset

As discussed in methodology Chapter 2, the INNO-Appraisal data collection process produced 2 different datasets which are shown in below Table 4. The second subset of the data, Sample B, only includes the templates pre-filled in by partners and then completed and verified by respective policy-makers. This sample is used to analyse section D onwards in the template which cover the issues that can only be known by policy-makers such as quality, usefulness, consequences, etc. The other dataset, Sample A, includes those templates covered by the Sample B and only-prefilled in templates for those evaluations there was no policy-maker response. Sample A is used to analyse basic characteristics of evaluations that were evident from the report. The overall number of observations in the Sample A and B are 171 and 132 respectively. Furthermore, Exhibit 5 depicts the composition of the Sample A.

Table 6: Sample Definitions

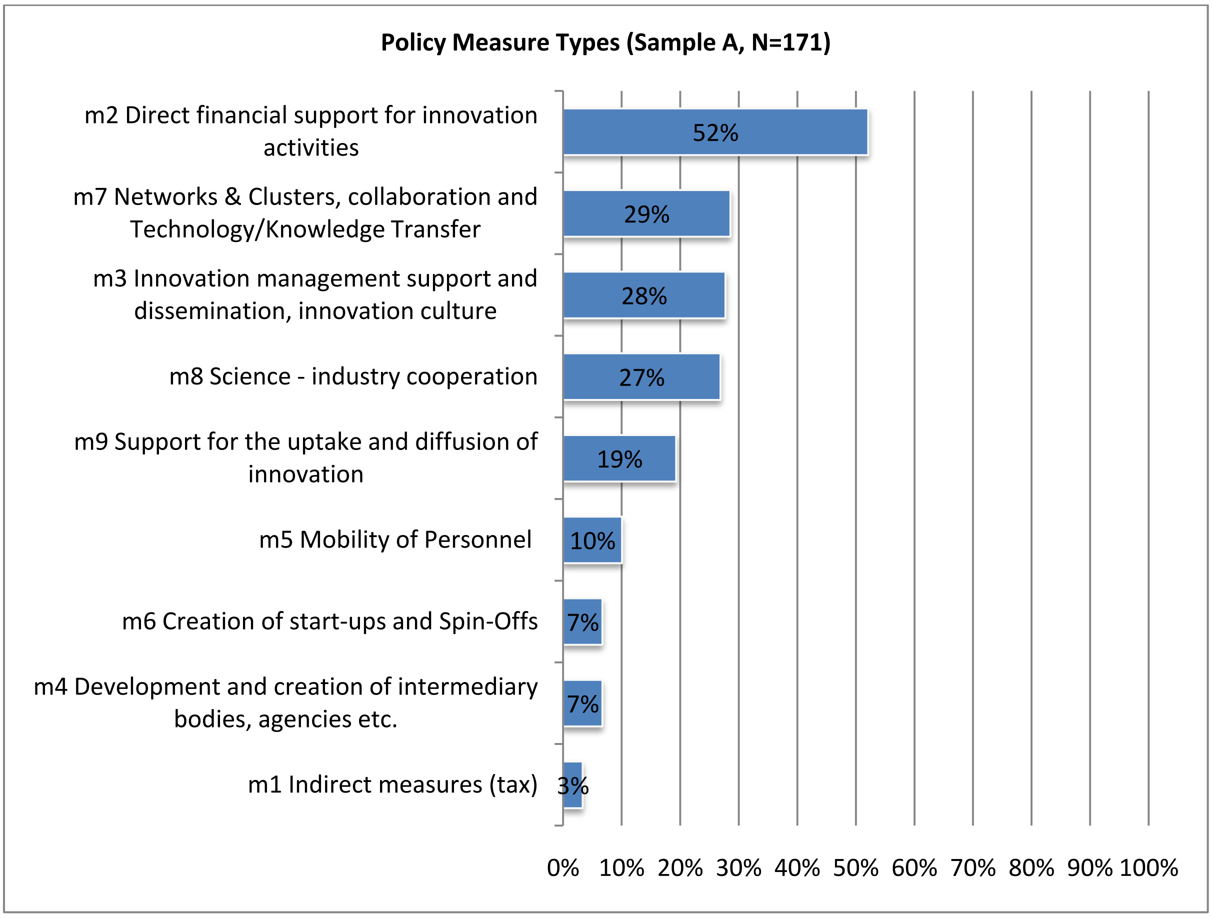
Sample	Definition	Potential use	n	remark
A	Includes all received questionnaires as well as the prefilled templates if no template was received;	Characteristics	171	GR, PT: since many individual evaluations are part of larger structural fund evaluations, Atlantis provided one 'umbrella' observation for each SF in order to avoid a country bias
B	Only received templates	Quality and Recommendations	132	GR, PT: individual templates of SF evaluations were used, because these were reviewed by different policy makers

Exhibit 5: The Type of Sources of Information for the Collected Data



As discussed in Chapter 2, as the last stage of the data collection after it appeared that INNO-Policy Trendchart could not be utilised to link the characteristics of policy measures to evaluations because of various problems, the INNO-Appraisal team analysed the associated policy measures to the evaluations in the database. Firstly, the team developed a rather simple typology based on the aim and modality of the measure. Within this typology, measures were assigned up to three of the categories shown Exhibit 6. This analysis revealed that the almost half of the evaluations in the INNO-Appraisal sample is for policy measures providing direct financial support for innovation activities. Circa 30% of evaluations are linked with measure types “networks & clusters, collaboration and technology/knowledge transfer”, “innovation management support and dissemination, innovation culture” and “science – industry cooperation”. Evaluations linked with “support for the uptake and diffusion of innovation” spans circa one-fifth of the sample while “mobility of personnel” type of measures constitutes 10%. “Development and creation of intermediary bodies” and “creation of start-ups and spin-offs” types of measures are both linked with 7% of the evaluations. Finally, only 3% of the evaluations in the INNO-Appraisal database are for “indirect measures (tax)”.

Exhibit 6: Types of Policy Measures Associated with Evaluations[8]

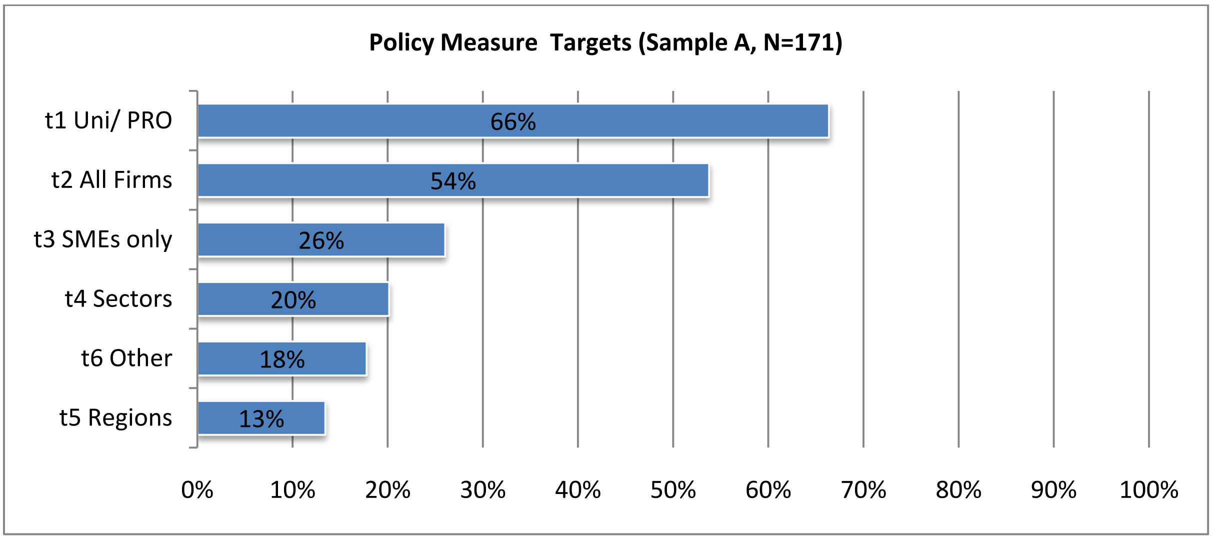


The team also characterised evaluations according to their respective policy measures' target groups. As depicted in Exhibit 3, two-thirds of the evaluations were for the measures targeting universities and/or public research organisations while this ratio is 54% for firms, 26% for only SMEs, 20% for sectors, 18% for other groups and 13% regions. The relationship between various evaluation characteristics and the type and target group of associated policy measures are explored more in depth later in this Chapter.

---

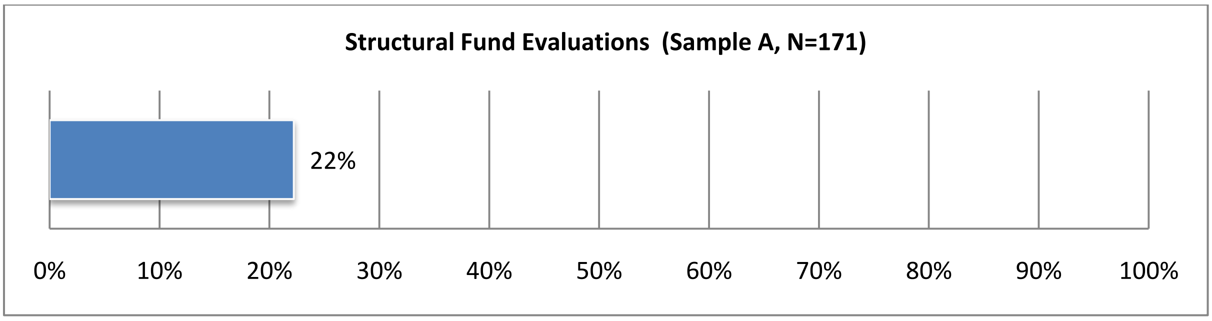
[8] Percentage of "Yes" response in Sample A, multiple response set

Exhibit 7: Target Groups of Policy Measures Associated with Evaluations[9]



Structural Funds evaluations comprise a significant part of our sample (roughly 22%), especially for Mediterranean countries and new-member states as depicted in Exhibit 8. Similarly, as shown in Exhibit 9, portfolio evaluations, evaluations that covered more than one measure in a single report, are significant (circa 13%). Around two-third of the sample is non-portfolio, non-structural-fund evaluations (Exhibit 10).

Exhibit 8: The Share of Structural Fund Evaluations[10]



---

[9] Percentage of "Yes" response in Sample A, multiple response set

[10] Percentage of "Yes" response in Sample A

Exhibit 9: The Share of Portfolio Evaluations[11]

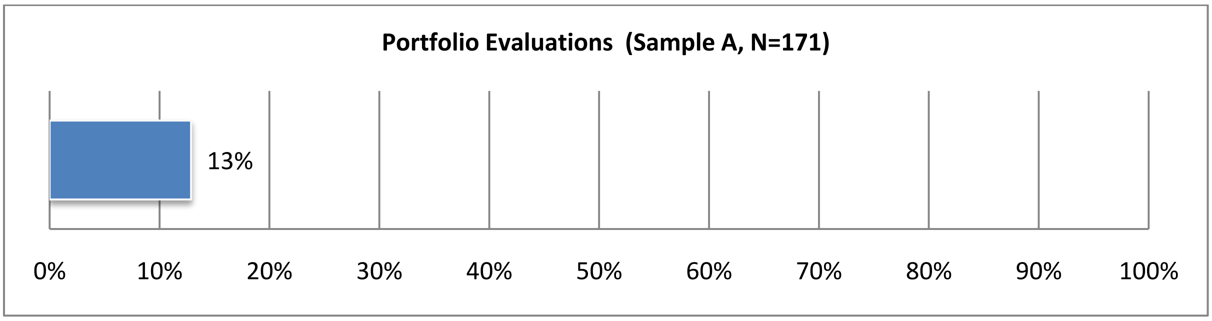
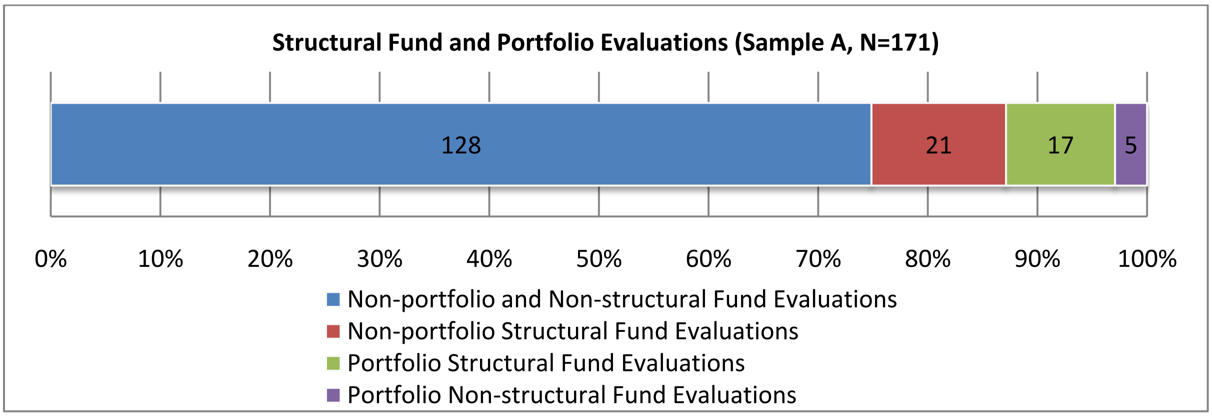


Exhibit 10: Composition of Structural Fund and Portfolio Evaluations[12]



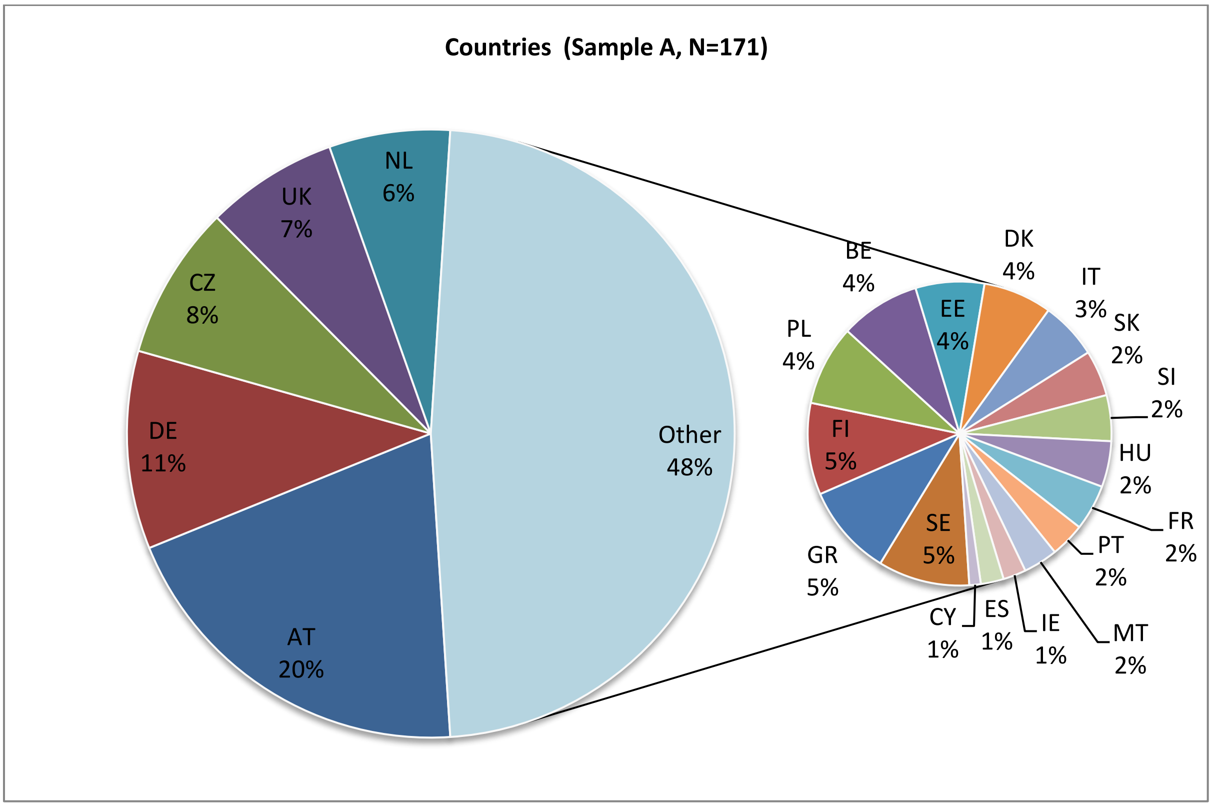
Austria is the leading country in our sample according to the number of evaluations. It has 34 evaluations which comprises roughly one-fifth of the sample. Other significant countries are Germany (11%), Czech Republic (8%), the UK (7%) and the Netherlands (6%). While half of the evaluations belong to these countries, the other half is shared by 17 countries. The overall distribution is given in Exhibit 11.

---

[11] Percentage of "Yes" response in Sample A

[12] Percentage of "Yes" response in Sample A, data labels show counts

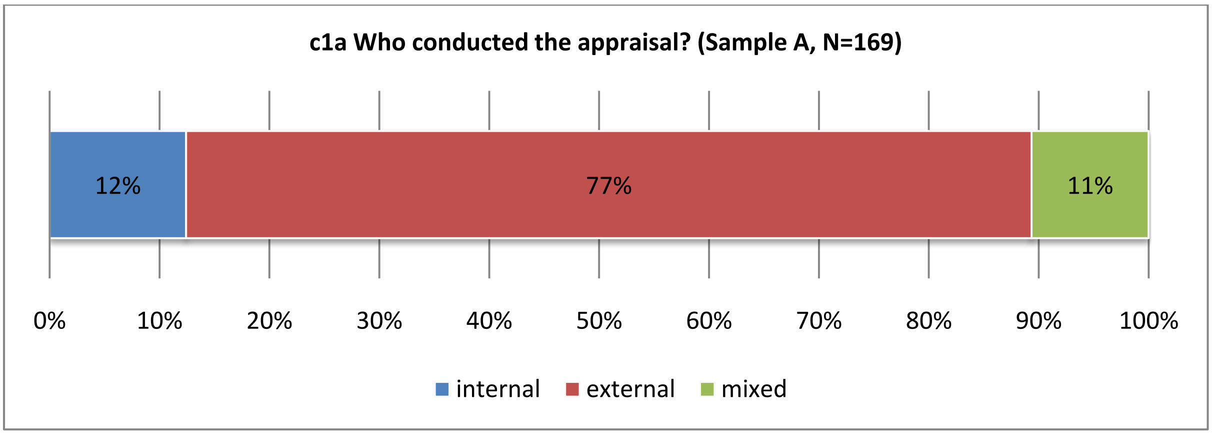
Exhibit 11: Country Distribution in the INNO-Appraisal dataset[13]



2.2 Evaluation characteristics

Most of the evaluators are external (circa 77%) as shown in Exhibit 12. The sample also contains around 12% of internally conducted and 11% of both internally and externally conducted evaluations.

Exhibit 12: Type of Evaluator

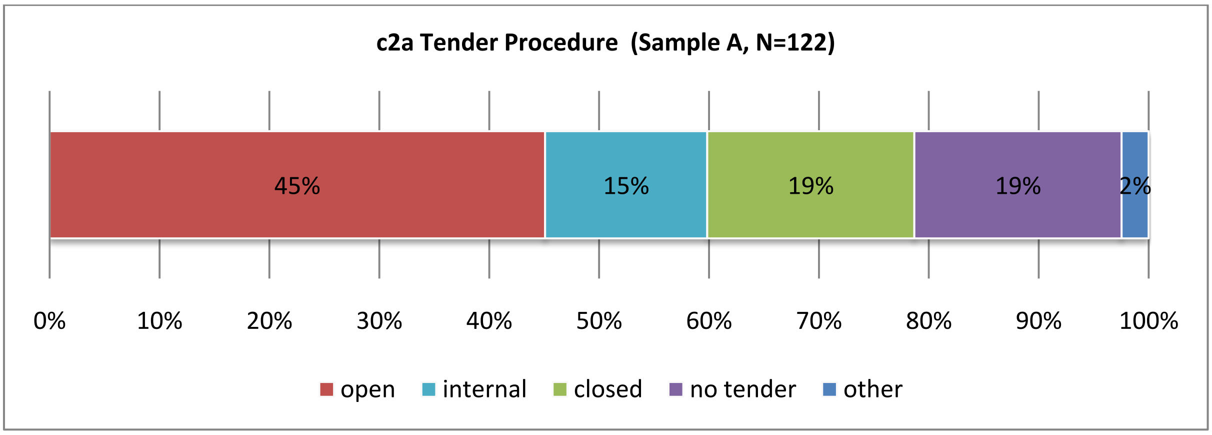


---

[13] As explained above, this distribution is no clear cut reflection of innovation policy evaluation culture in the countries, but reflects the reporting of innovation policy and evaluation in the Trendchart database. Because of different ways of reporting, e.g. Finland having many evaluations not as among their Trendchart measures there are activities of innovation agencies bundled but not represented individually on the first entry level at Trendchart, leading to a severe underreporting for Finland.

Exhibit 13 shows that around 45% of evaluations tendered in an open procedure, while internal, closed and non-tender procedures are each about 15%, 19% and 19% respectively. The N value of 122 is lower than the overall N of 171, however, due to the tacit nature of this information.

Exhibit 13: Tender Procedures of Evaluations



Ex-ante and interim evaluations each comprised around 15% of our sample, while the accompanying evaluations are around 43% as shown in Exhibit 14. Finally, ex-post evaluations formed around 30% of our sample. There is a risk that the sample has a bias against ex-post evaluations as the data collection was based on the INNO-Policy Trendchart database which only included running measures at the time of data collection and this led to the exclusion of ended measures and thus their ex-post evaluations.

Exhibit 14: Timings of Evaluations

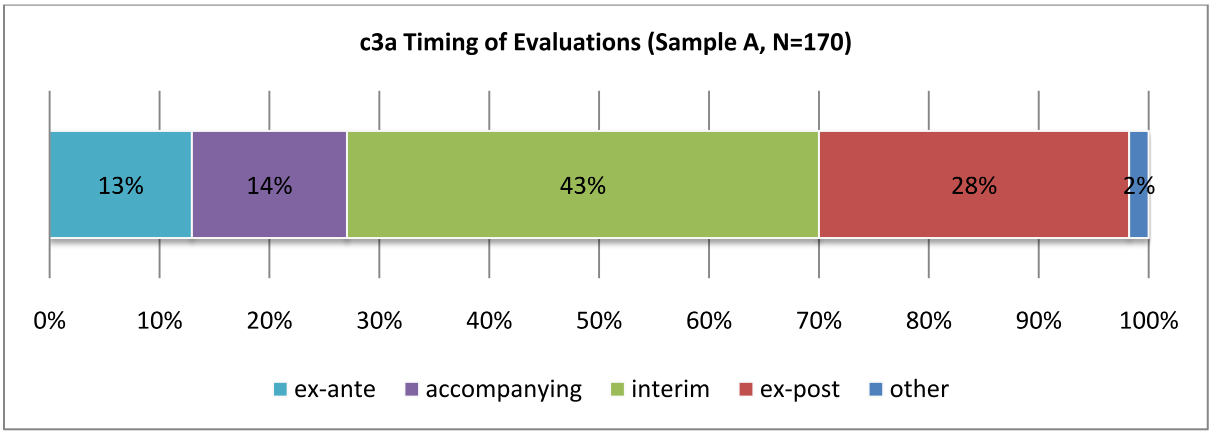


Exhibit 15 depicts that summative evaluations are over 20% while around two-fifth of the evaluations are formative. Finally around one-third of evaluations exhibited both summative and formative characteristics.

Exhibit 15: Purpose of Evaluations

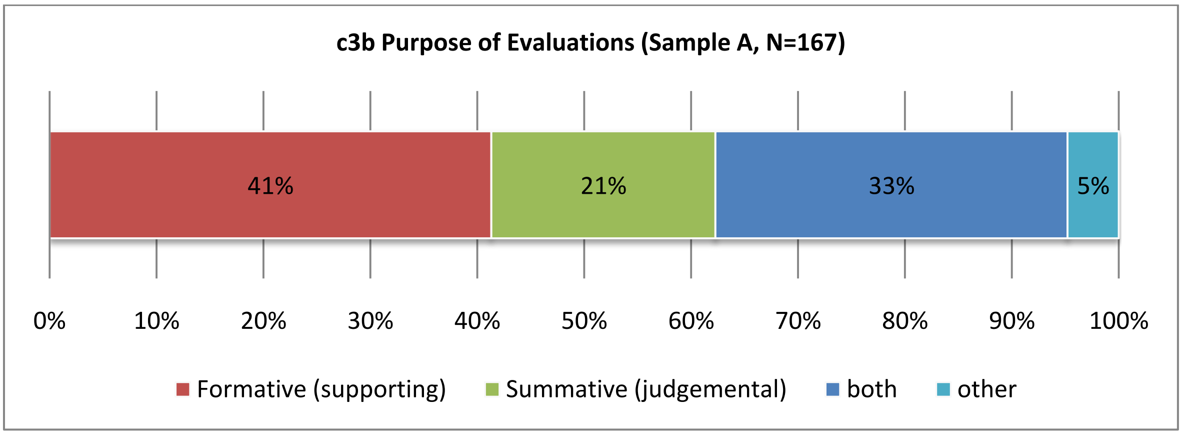
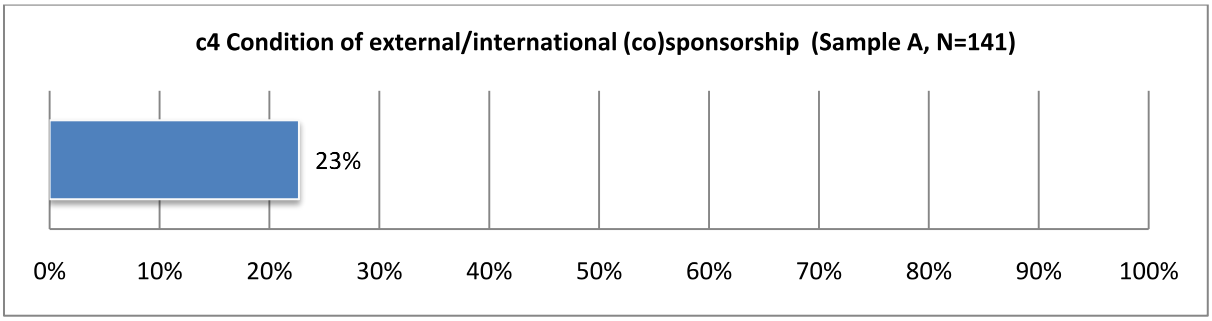


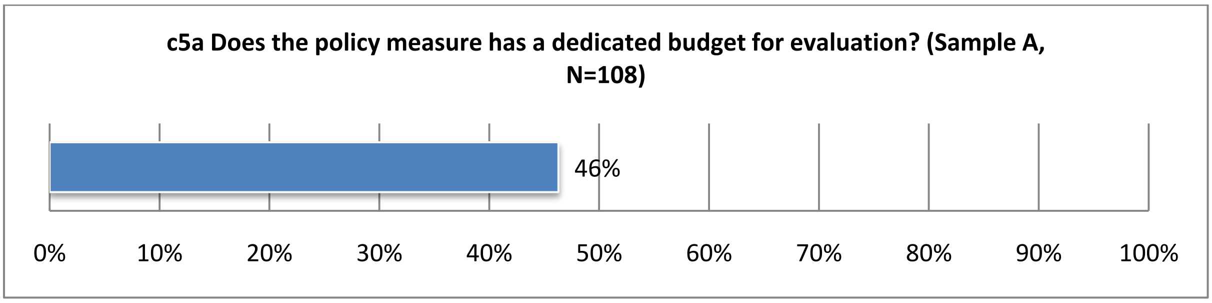
Exhibit 16 shows that a quarter of the evaluations in our sample were conducted as a condition of external/international (co)sponsorship. The high number of missing values for this element should also be noted.

Exhibit 16: Evaluations Sponsored Externally[14]



Around 46% of evaluations in our sample had a dedicated budget. There are more than 40% missing values for this question as depicted in Exhibit 17, therefore, it is highly likely that the share of evaluations with dedicated budgets is higher than 50%. Exhibit 18 shows that 67% of the evaluations were planned for while the rest were not.

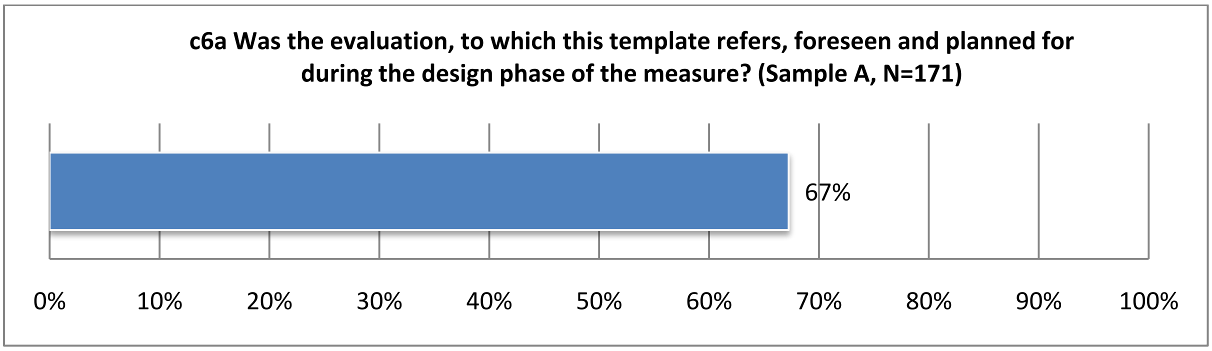
Exhibit 17: Evaluation Budget



---

[14] Percentage of "Yes" response in Sample A

Exhibit 18: Evaluation Planning



The very general tropic “outputs, outcomes, impacts” is the evaluation topic most often covered, with nine in ten appraisals covering this topic as shown in Exhibit 19. The other significant topics are “goal attainment / effectiveness” (circa 89%), “internal and external consistency” (both circa 80%), “policy / strategy development” and “programme implementation efficiency” (both circa 76%). “Coherence / complementarity” was covered by 62% of evaluations while this ratio is 50% for “quality of outputs”, around 50% for input, output and behavioural additionality and 47% for “project implementation efficiency”. The least frequent evaluation topics are “gender issues” (circa 24%), value for money (circa 27%) and finally “minority issues (circa 7%).

Exhibit 19: Evaluation Topics[15]

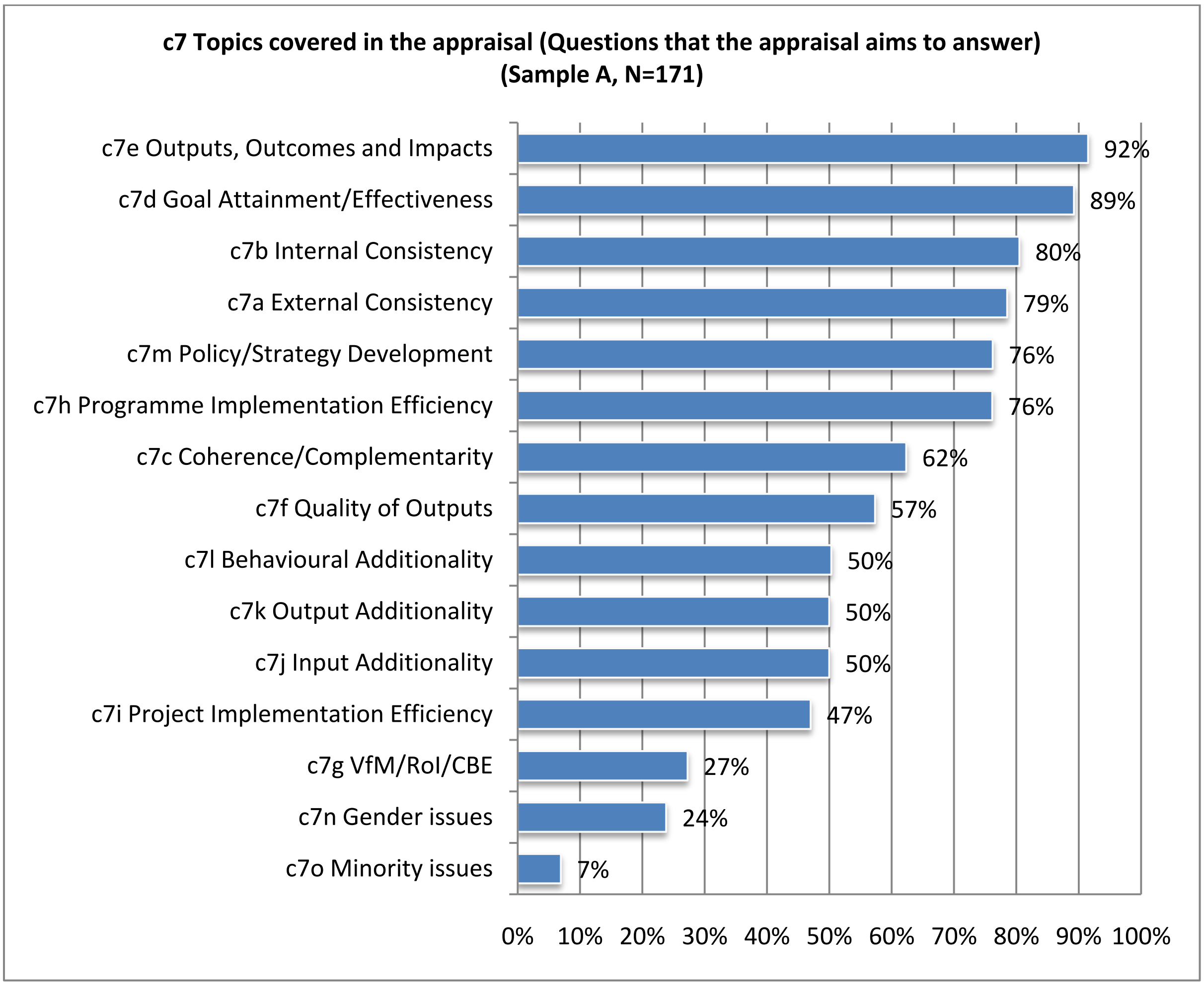


Exhibit 20 depicts how topics cluster in evaluations. According to this figure, the closest link is between “goal attainment / effectiveness” and “outputs, outcomes, impacts”; if one of them covered in an evaluation the other one is highly likely to be covered as well. The second cluster is comprised of “internal” and “external consistency”. These four topic form a big cluster together and consequently linked with “policy / strategy development”, “programme implementation efficiency” and “coherence / complementarity”. These topics form a big cluster that can be considered as the core topics of evaluation. In fact, they are the most popular topics and covered in most of the evaluations (Exhibit 19). The second big cluster is formed by three closely related variants of additionality, “quality of outputs” and “project implementation efficiency”. Value for money forms another big cluster on its own. Finally, fringe topics of “gender and minority” issues compose a final group.

---

[15] Percentage of “Yes” response in Sample A, multiple response set

Exhibit 20: Clustering of Topics[16]

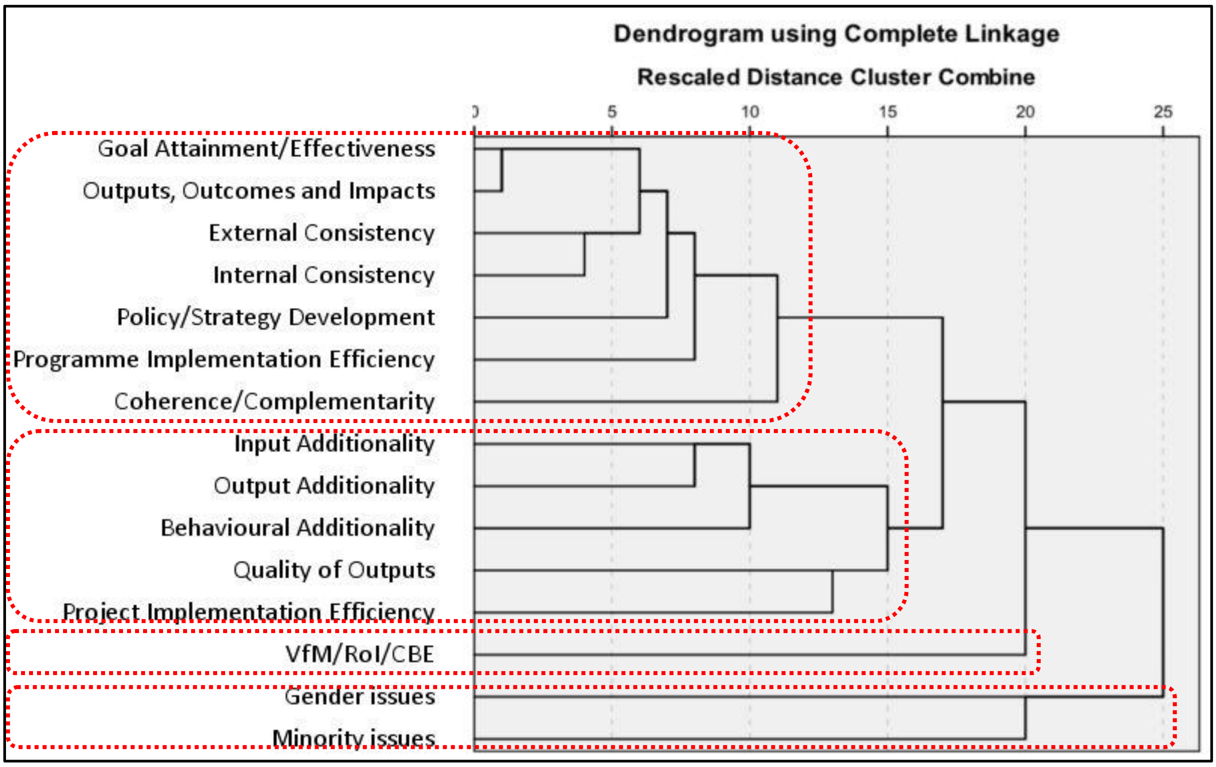
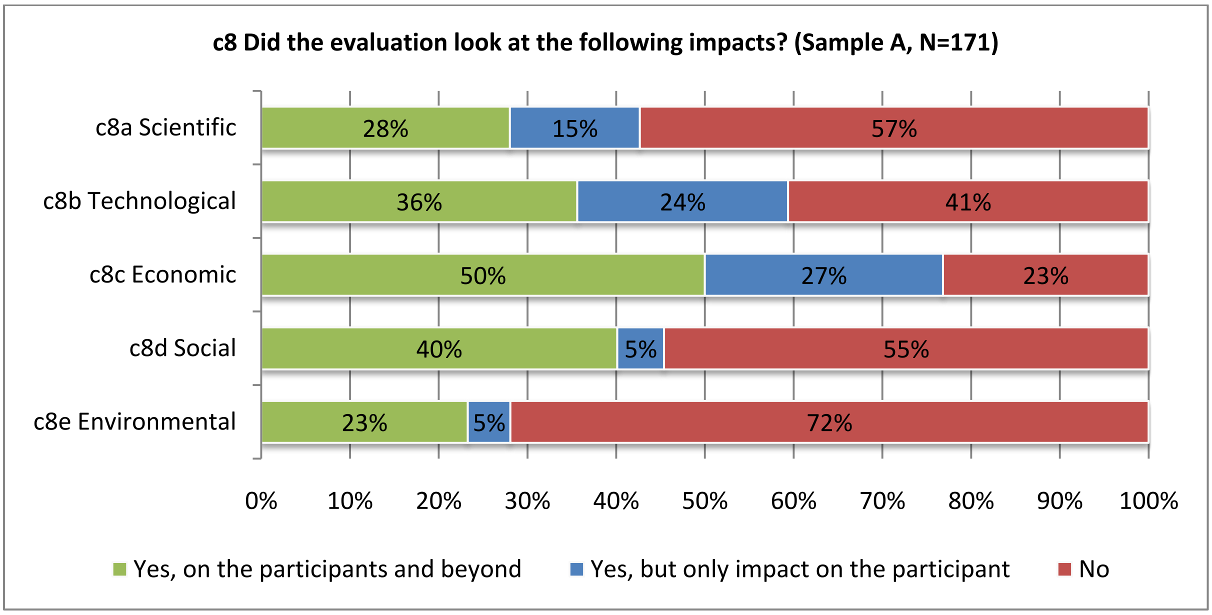

Dendrogram using Complete Linkage
Rescaled Distance Cluster Combine
0
5
10
15
20
25
Goal Attainment/Effectiveness
Outputs, Outcomes and Impacts
External Consistency
Internal Consistency
Policy/Strategy Development
Programme Implementation Efficiency
Coherence/Complementarity
Input Additionality
Output Additionality
Behavioural Additionality
Quality of Outputs
Project Implementation Efficiency
VfM/RoI/CBE
Gender issues
Minority issues

Exhibit 21 presents how different kinds of impacts are covered, whereby for all of five impacts asked for the template differentiated for impacts on "programme participants only" and "beyond participants". It shows that the impact most often covered in evaluation of innovation policy measures is economic (77%), followed by technological impacts (60% of evaluations). Social and scientific impacts are both equally frequent (45% and 43%), while 28% of the evaluations examined environmental impacts. What is remarkable is that for all impact dimensions the share of evaluations which claim to look beyond the participants also is higher than those that are limited to on the participants. This appears to reflect the growing need to demonstrate for societal and broader economic benefit of polices.

---

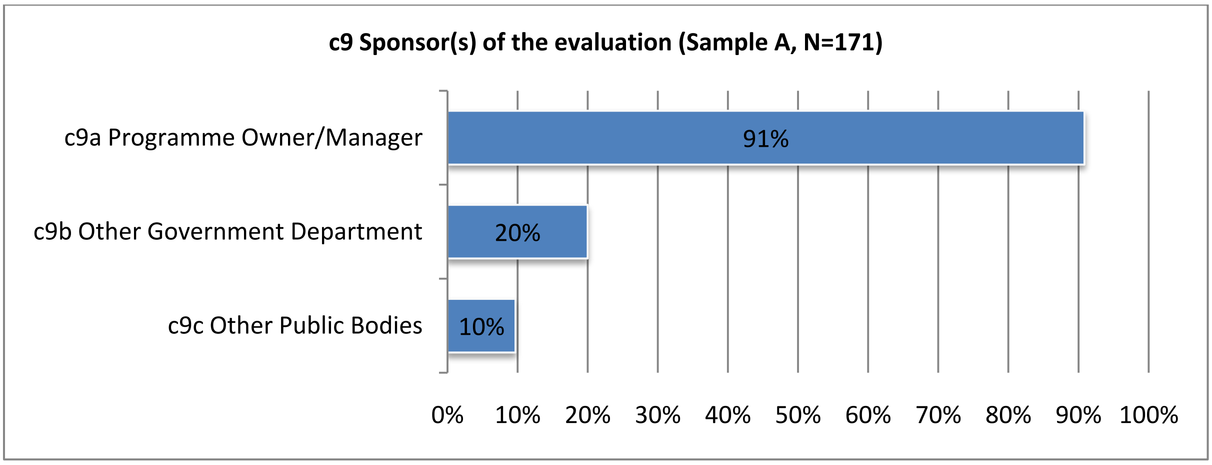
[16] Rezankova (2009) recommends "Jaccard's co-efficient" or "Yule's Q" measures for object clustering (clustering of variables of same type) of dichotomous (variables that take binary options) asymmetric ("1" and "0" values are of inherently different importance) variables. This method does not cluster variables on the basis of co-absence of same trait (i.e. both variables takes the value "0" at the same time). In this analysis, furthest neighbour method which links topics with complete linkage is used by applying Jaccard's co-efficient measure.

Exhibit 21: Impacts Looked at in Evaluations[17]



Evaluations are pre-dominantly commissioned by programme owners / managers themselves, which are commission 91% of all evaluations in the dataset (Exhibit 22). Sponsorships of other government departments and other public bodies do not constitute a significant share in our sample with about 20% and 10% respectively.

Exhibit 22: Evaluation Sponsors[18]



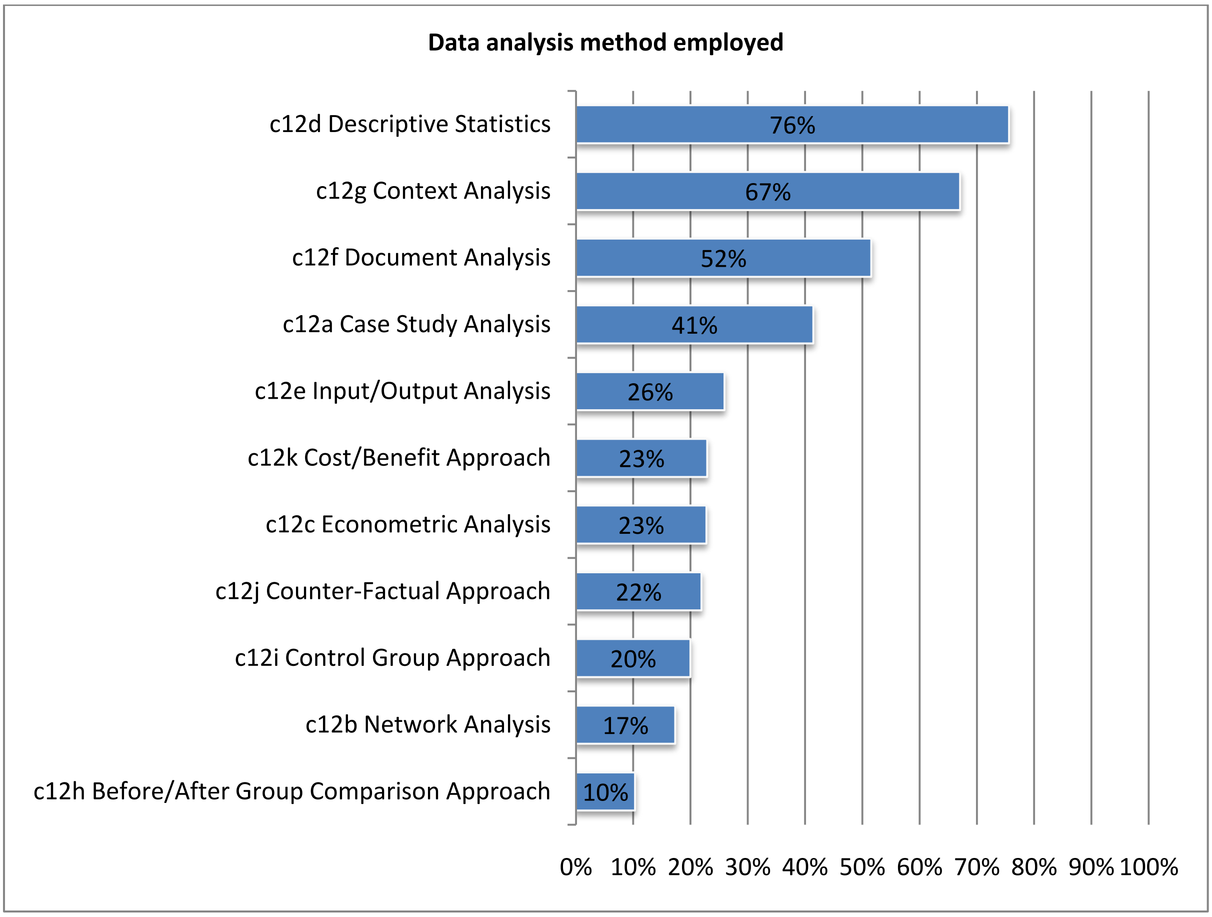
As regards methodology, the template asked for a broad range of methods. Almost three quarters of the evaluations in our sample employed descriptive statistics, while two-thirds employed context analysis. Document analysis was employed in 52% of appraisals. This is followed by case study analysis (41%), input / output analysis (around 26%), econometric analysis (around 23%), and network analysis (around 17%) as depicted in Exhibit 23. Furthermore, within that question in the template, certain evaluation designs and approaches were considered. In consequence, it is found that around a quarter to one fifth of evaluations employed “cost-benefit approach”, “counter-factual approach” or control group approach”, while 10% used “before/after group comparison approach” as their main approach/design (Exhibit 23).

---

[17] Percentage of "Yes" response in Sample A, multiple response set

[18] Percentage of "Yes" response in Sample A, multiple response set

Exhibit 23: Data Analysis Methods and Main Evaluation Designs/Approaches Employed in Evaluations[19]

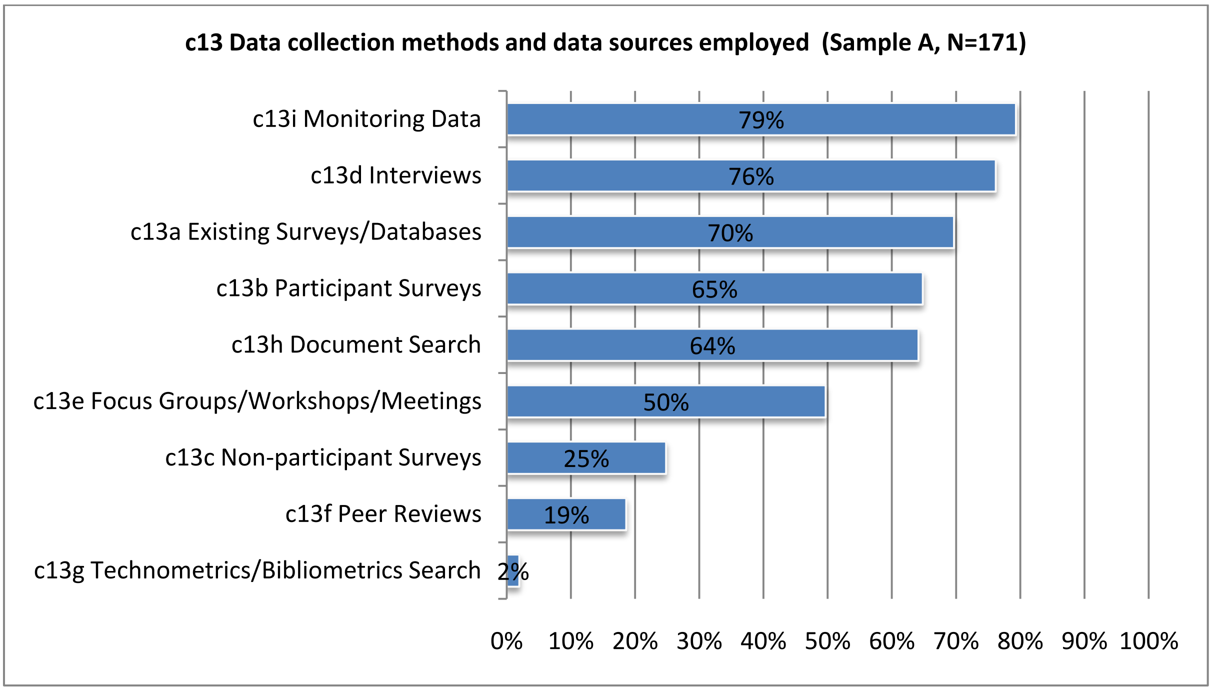


Next to methodological approaches the template further asked for data collection methods Monitoring data is the most popular data collection method (circa four-fifth), followed by “interviews” (circa three-quarters), “existing surveys / databases” (70%), “document search” and “participant surveys” (both around three-quarters) and “focus groups / workshops / meetings” (around half) as shown in Exhibit 24. “Non-participant surveys”, “peer reviews” and “technometrics / bibliometrics search” are all used in less than a quarter of evaluations in our sample.

---

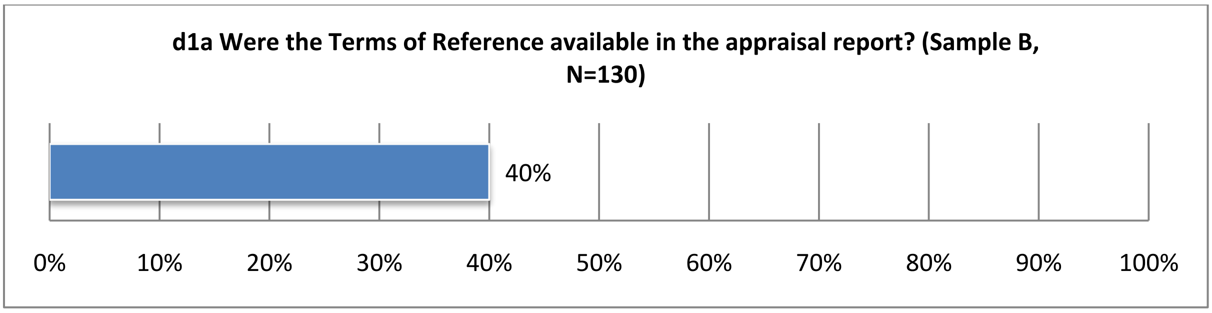
[19] Percentage of “Yes” response in Sample A, multiple response set

Exhibit 24: Data Collection Methods and Data Sources Employed in Evaluations



An important aspect of quality of the evaluation process is the commissioning process, the mode of the tender process. Exhibit 25 shows that nearly 40% of evaluations in the sample included their terms of reference and for more than half of the evaluations that did not include them in the report, terms of references were available in elsewhere (Exhibit 26). In almost all of these two categories of evaluations, terms of references clearly stated the objective of the appraisals while almost one-third of them specified the methodologies and approaches (Exhibit 27 and Exhibit 28).

Exhibit 25: Availability of Evaluations Terms of References as Part of Report[20]



---

[20] Percentage of "Yes" response in Sample B

Exhibit 26: Availability of Evaluations Terms of References in Other Sources[21]

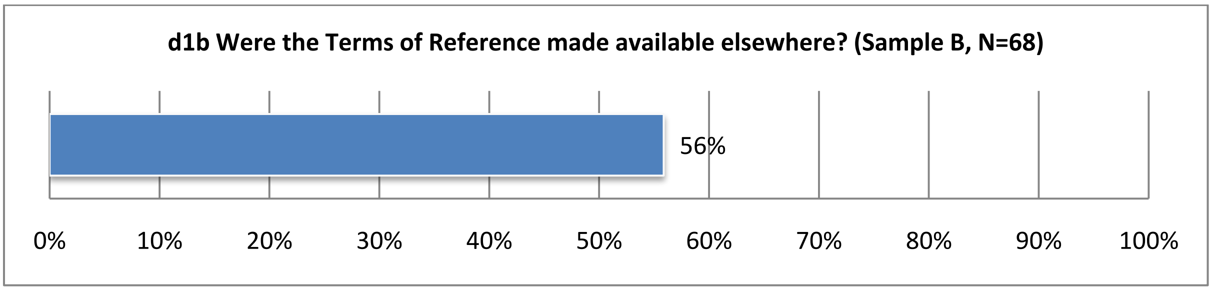


Exhibit 27: Clearly Stated Objectives in Evaluation Terms of References[22]

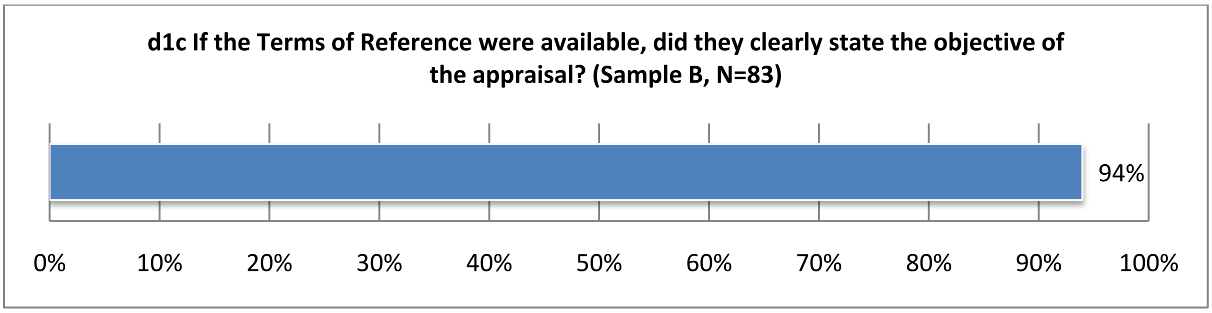
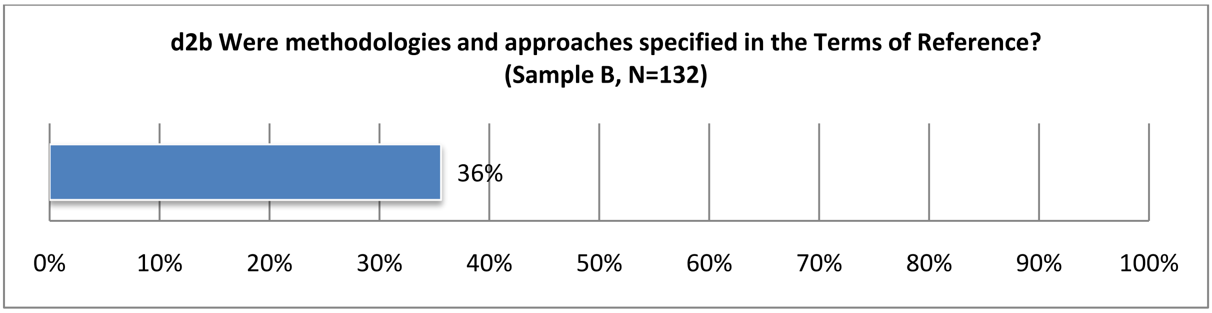


Exhibit 28: Specification of Methodologies and Approaches in Terms of References[23]



As expected and as shown in Exhibit 29, most of the evaluations were produced for different categories of policy makers (programme owners and other related government officials). Still, a surprisingly high number of evaluations (two thirds) claim to target politicians as well, highest level leaders of ministries or parliamentarians. However, only half of the evaluations target the users, and 40% potential users, the potential to mobilise the community appears to be not fully exploited. Half of the evaluations are also produced for external auditors, which means that external accountability is as important as mobilisation. The general public plays a minor role as target group, which is – given the enormous number of rather limited and targeted measures are of limited interest to the general public, however, as with potential users, it appears that evaluations could be used more often to inform the public about direction and effects of innovation policy.

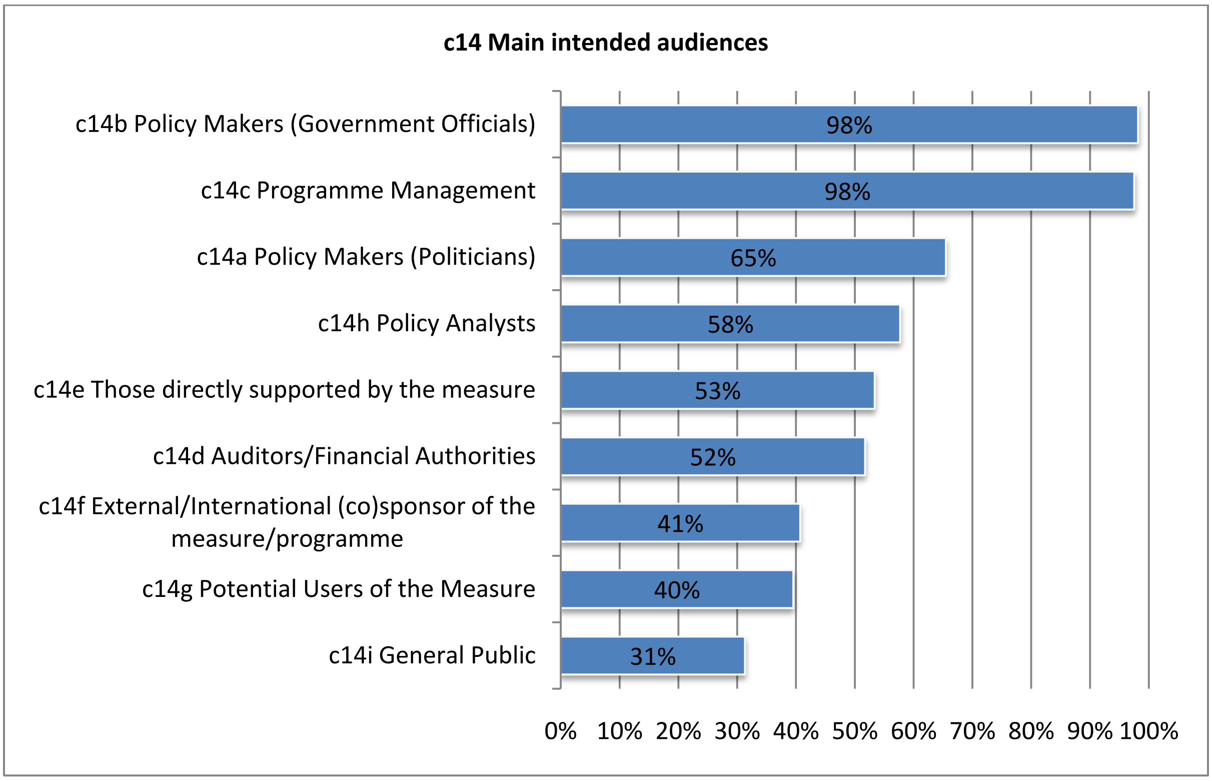
---

[21] Percentage of “Yes” response in Sample B

[22] Percentage of “Yes” response in Sample B

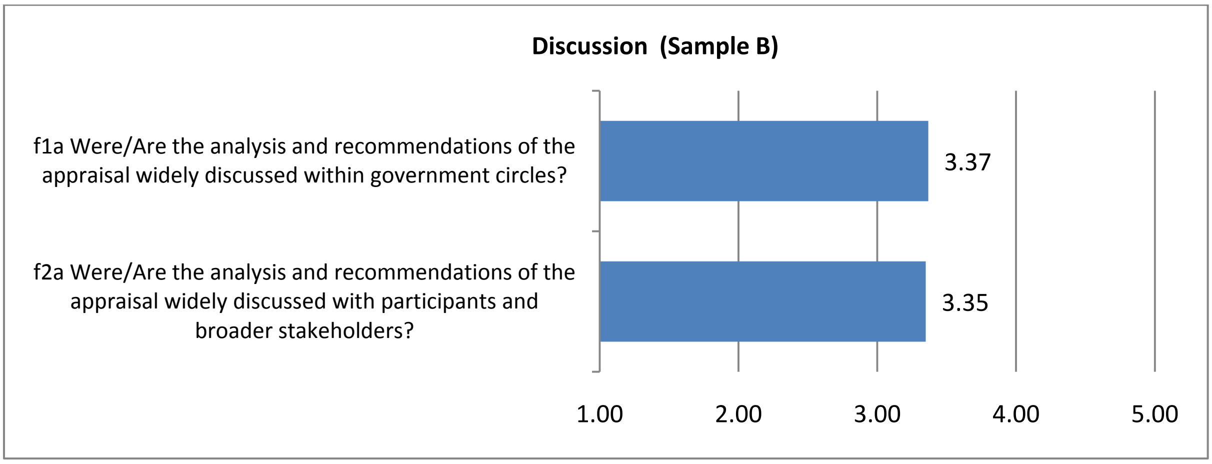
[23] Percentage of “Yes” response in Sample B

Exhibit 29: Main Intended Audiences of Evaluations[24]



The template further asked for the dissemination of evaluations. Exhibit 30 shows the perceived extent of the dissemination by using the usual Likert scale. As expected, although the results are quite close, evaluations in our sample are slligtly more discussed in governement circles than with participants and wider stakeholders, but in general the values on our scale from XX to XX appears moderate only.

Exhibit 30: Discussions of Evaluations[25]



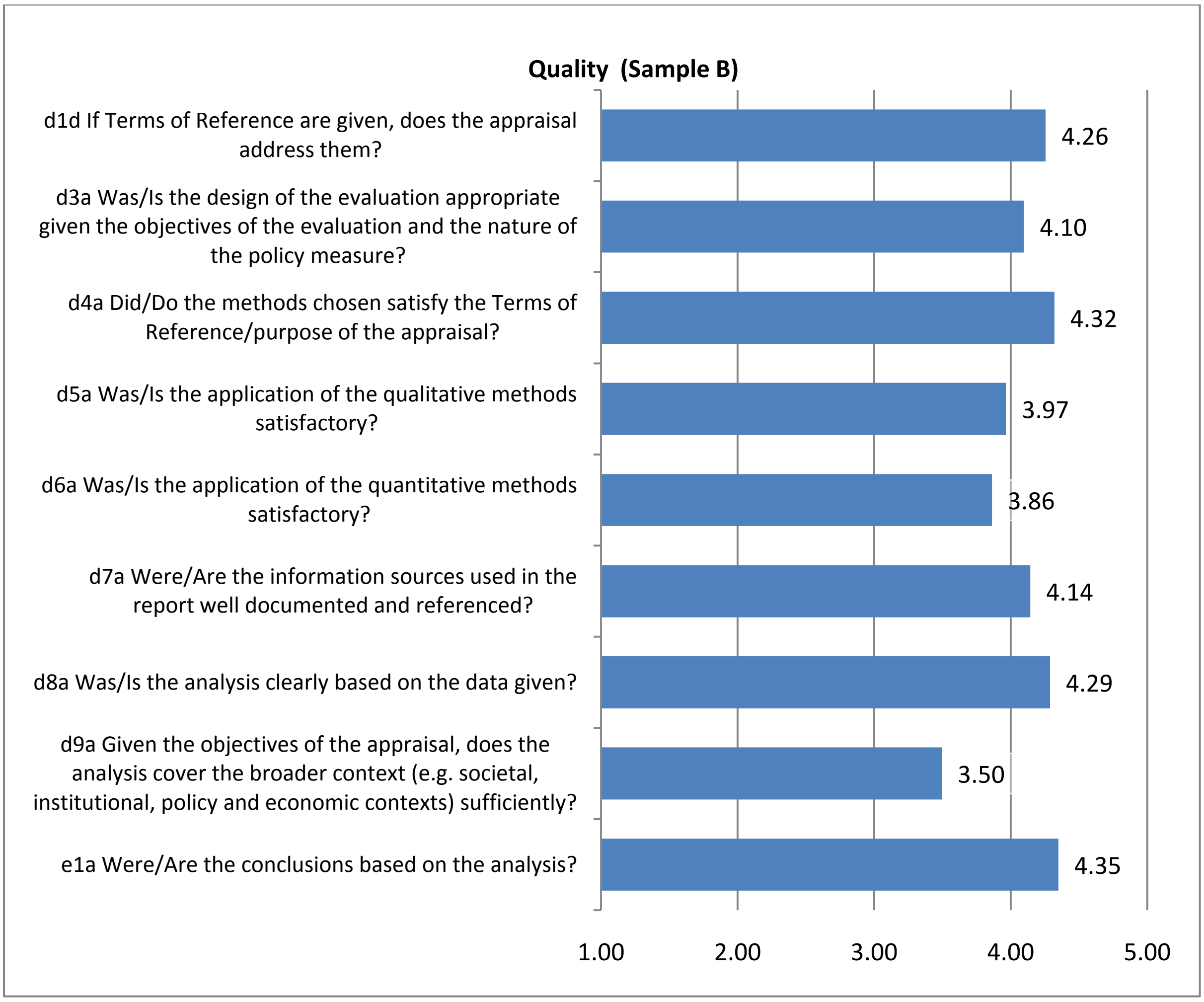
---

[24] Percentage of "Yes" response in Sample A, multiple response set

[25] 1-5 Likert scale where only edge values specified (1 is "no, not at all" while 5 is "yes, definetly"), Sample B, means used

The template also asked for the perceived quality of evaluations (Exhibit 31). This part of the template was assessed by the respective policy makers. For each quality variable, a 1-5 Likert scale is used where only edge values specified (1 is “no, not at all” while 5 is “yes, definetly”). In general the perceived quality of evaluations in our sample is quite high ranging between the highest item of “conclusions based on the analysis” with an average of 4.35 and the lowest item of “analysis cover the broader context sufficiently” with and average of 3.50.

Exhibit 31: Perceived Quality of Evaluations[26]



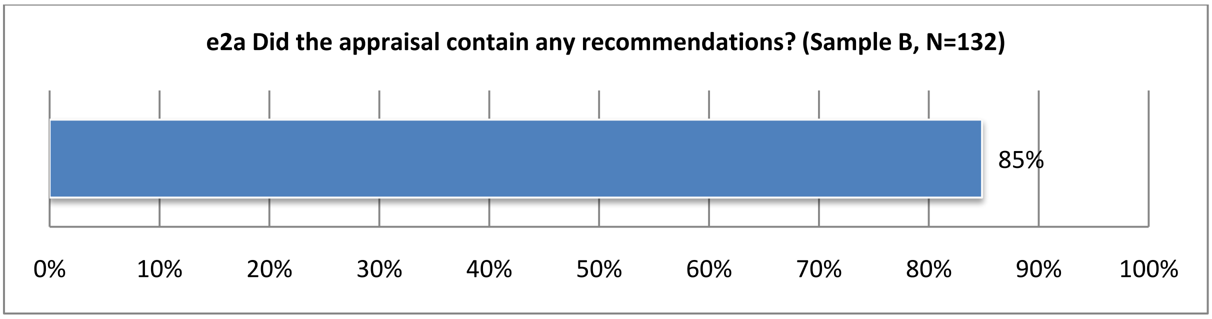
To get a more general appreciation of the overall quality of the evaluation reports, a new compositve variable for high-quality evaluations was created manually. Based on the expertise of the research team, those evaluations are considered of ‘high quality evaluations’ that score high (>3 on the Likert scale (see Exhibit 31 above) on *each of the following major quality variables* of an evaluation; namely: appropriate design (d3a), analysis clearly based on given data (d8a) and conclusions based on analysis (e1a). In total 81 out of 132 evaluations (61%) meet these conditions and thus can be regarded as high quality.

---

[26] 1-5 Likert scale where only edge values specified (1 is “no, not at all” while 5 is “yes, definetly”), Sample B , means used

Evaluations are done to improve policy, and thus one would expect that they include recommendations on the basis of their analysis. Exhibit 32 shows that three quarters of the evaluations in our sample included recommendations.

Exhibit 32: Recommendations in Evaluations[27]

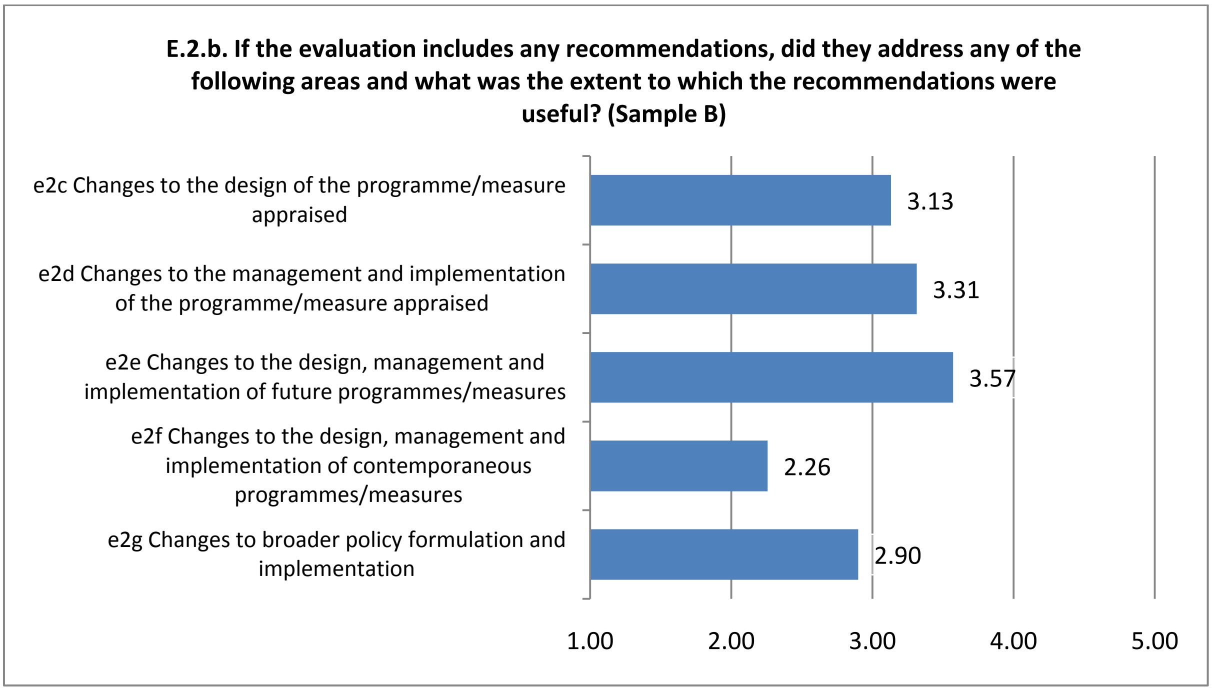


As can be seen from Exhibit 33, however, these recommendations are percieved as being of moderate usfulness. The question asekd in the template was if the recommendations covered a set of issues and if so, how useful the recommendations were. Again the same Likert scale used for the perceived quality dimension is employed here. Usefulness was ratest highest for management and implementation of measure and the lowest in the design of contemporaneous programmes. As usefulness of recommendations and of evaluations more generally is of utmost importance – the essence of doing evaluations in the first place – it will be analysed more in depth in a case study in a further chapter of the report.

---

[27] Percentage of “Yes” response in Sample B

Exhibit 33: Perceived Usefulness of Recommendations of Evaluations[28]

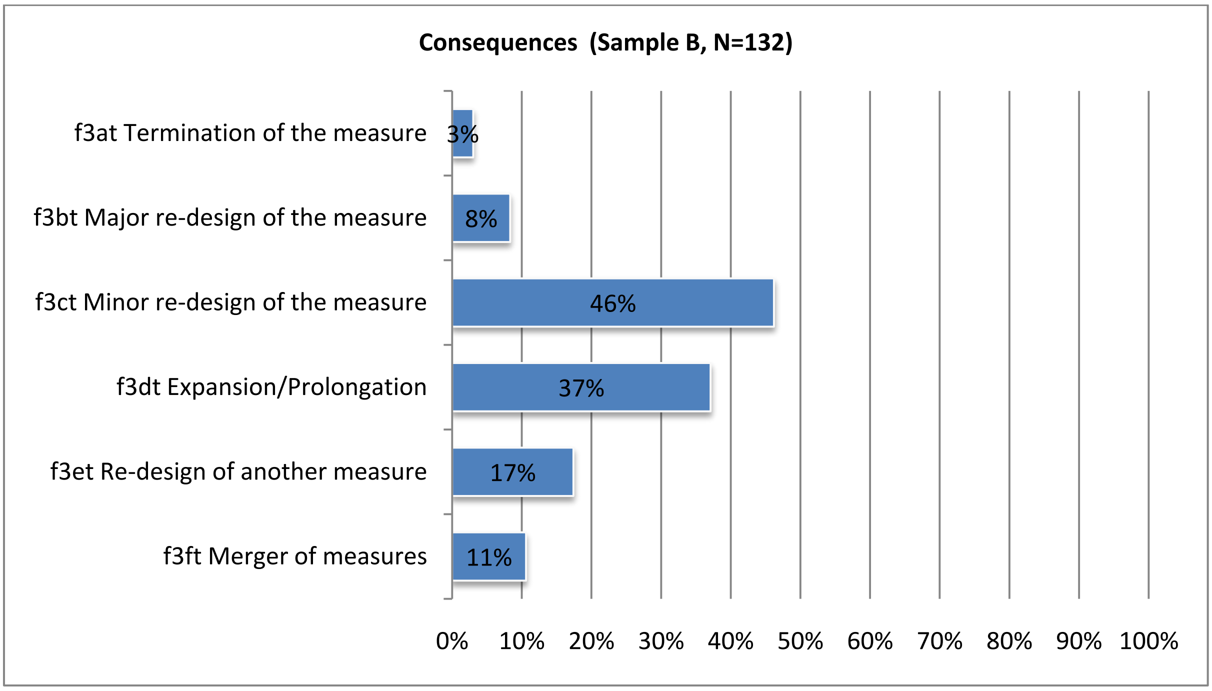


Finally, the template inquired if the evaluations led to certain consequences. It offered seven different consequences (Exhibit 34). The three most often mentinoed consequences of evaluations are “minor re-design of the measure” (around 46%), “expansion / prolongation of the measure” (37%) and “re-design of another measure” (around 17%), the latter pointing to a braoder learning effect of evaluations.

---

[28] 1-5 Likert scale where only edge values specified (1 is “no, not at all” while 5 is “yes, definetly”), Sample B, means used

Exhibit 34: Consequences of Evaluations[29]



The following section will analyse the evaluations in more detail, linking various evaluation characteristics through cross-tabulation and correlation.

3 Exploring and Understanding Evaluations – some cross-cutting analysis

The previous section has presented an overall picture of the practice of evaluation in innovation policy in Europe. On that basis, a next step in the analysis is to explore and understand more in-depth what determines the design and effects of evaluations. The variables to be used for this analysis are all characteristics of the evaluations (as collected in the templates) and the policy measure categorisation (type of measure, target group).

The in-depth questions that were asked in the template offer a potentially very large number of possible relationships. The selection presented in this section is the result of a broad screening and discussion process and it reflects those relations that are most relevant for the practical knowledge as to evaluation practice in Europe. Some of the most important issues will only be reported shortly in this section, but then deepened in our case studies in section 4 onwards. The following section starts with the characterisation of evaluations for the different policy measures and target groups, before section 3.2 presents the relation of evaluation characteristics. Subsequently, section 3.3 takes an in-depth look at the determinants of quality and consequence out of evaluations, before a last section undertakes the attempt to explore if the large number of evaluations can be grouped into clusters.

---

[29] Percentage of "Yes" response in Sample A, multiple response set

We need to stress again that the interpretation of the data needs to be made with caution. The relations we interpret are correlations, not causalities. As with all correlations, we need to stress that those correlations do not suggest simple or even uni-directional causality, but rather a systematic link of criteria that nevertheless tell us something about the nature of evaluations.

3.1 The meaning of policy measures

During the life-time of INNO-Appraisal, one key question of those concerned with designing and implementing innovation policy and related evaluation was if there is a systematic link between the types of measures and the nature of the evaluations used. Thus, this section will relate some key characteristics of the evaluations in our database to the policy measure type, the distribution of which was presented in previous sections. It will not interpret all relations we have tested nor will it show all of the results, but focus on the most significant and clear findings.

A second, related issue is to look at the relation between the target group of measures and evaluations. The assumption here would be that different kinds of target groups have different expectations, support needs and backgrounds, that would have to be taken into consideration by evaluation design. However, the statistical analysis along the dimension 'target group' did not result in many meaningful systematic relations, it appears that the measure type is more significant than the target group.- This is consistent with the fact that many measures in fact have more than one target group which blurs the statistical analysis. In the following we will only report results for the target group dimension in exceptional, clear and relevant cases.

A first major finding of the policy measure related analysis of evaluation practice in Europe is that there are a couple of evaluation dimensions for which policy measure type does not make a difference, which means that for those issues other considerations are more important than the type of measure.

Thus, the variation for the following issues is not significantly influenced by the policy measure type:

- tender process,
- the use of internal vs. external evaluators,
- the use of formative vs. summative evaluations
- having a dedicated budget or having the evaluation planned in the first place
- having external sponsor of the evaluation or the programme (in other words, the type of measure does not make a difference as to the likelihood of having an external sponsor of the evaluation)
- coverage of impact, outcome and output (as most evaluations cover those aspects)
- coverage of programme implementation efficiency
- coverage of policy and strategy issues
- coverage of gender and minority issues
- the types of audiences (which is interesting, as intuitively one would expect that diffusion oriented programmes are more likely to be geared towards a more general public).
- use of data sources and methods. Overall, the use of data sources and methods does not greatly differ for the different types of programmes – even if some differences remain (see below). In combination this means that evaluations in terms of their overall approach appear

to be rather standardised, using a set of methods and data sources, there is no strong specific bias towards methods and sources for different kinds of programmes.

This list is highly relevant. It shows that other factors, such as organisational and country specific traditions and general practices dominate the design and implementation of evaluations to a large extent, not so much the evaluation object – the policy measure – itself.

However, a set of interesting specific characteristics of evaluations of different policy measure types are observed. As for evaluation topics and impacts that are covered in evaluation, the following observations stick out (as statistically significant):

- In contrast to intuition, **complex cooperation programmes** are not more likely to be evaluated for internal and external consistency. This is interesting as the linkage of actors by definition means to link different social, economic, technological etc. contexts, and one would expect this to be reflected in a more systematic consistency check
- Evaluations in **networking programme** are more likely to look for efficiency on project level, while **broad, diffusion oriented programmes** are less likely to be concerned with project efficiency issues
- Technological and scientific impacts are both slightly more common in **indirect support measures, management / culture measures** and **cluster/network measures**. In addition, scientific impact is – as top be expected – more often covered in **industry-science collaboration**.
- Societal impact in evaluations is only positively correlated with **science – industry collaboration** measures, not, however, with **diffusion measures**. This, one can expect, is an expression of the horizontal nature of innovation policy (as opposed to the theme oriented nature of science and science – collaboration measures), whereby innovation diffusion measures are not linked with certain societal goals. The current move towards challenges and missions might lead to a change here, whereby innovation effects and societal effects might be much more coupled in the future.
- Economic impact is more often covered in **direct subsidy measures** and in **cluster / network measures**, as both of those programme types are most directly linked to economic output.

In terms of methods and data, beyond the general uniformity of approaches between different measure types, we find that

- Network analysis is consistently used more in all programme types that link actors
- Interviews and focus groups/workshops are more often used in clustering and networking programmes as well as in direct subsidy programmes to support innovation activities

There are some differences as regards the target groups of evaluation:

- One important finding is that for **cooperation and networking programme** the beneficiaries are also an important audience of the evaluations, i.e. The need for learning is reflected in the way evaluations are discussed, as input for constant improvement.
- International sponsors of programmes are more likely to be an audience for the evaluation in **innovation management programmes**. This is consistent with the purpose of many structural fund measures that have a focus to build up competencies in industry.

- Evaluations in **complex cooperation programmes** target beneficiaries and potential beneficiaries, as those are knowledgeable about the cost-benefit ratios. To overcome hesitations they have because of transaction and learning costs, they need re-assurance as to why the programmes is sensible for them
- The general public as such is not an important target group, and there are no specific types of measures for which the general public would be an important audience.

Very interestingly, policy makers responsible for measures that **support networking, clustering or concrete science-industry cooperation** tend to be more prescriptive when it comes to methods used in the evaluations (ToR specifying methods). It appears that they are eager to get a specific dimension (benefit out of networking and cooperation) well covered and they link this with the use of certain methodologies. This is confirmed by interviews around behavioural additionality, more concretely, the cooperation aspect of it, where methods have been discussed widely by and with policy makers.

In terms of the likelihood of consequences out of evaluations, there are no systematic differences for different kinds of measures, with the only exception being prolongation of measures as an evaluation consequence. This is more likely in evaluations of **cluster and networking programme**, where there appears to be a greater need for actually showing effects.

The evaluations are perceived (by policy makers) as being of quite different quality, whereby quality is defined by a range of different quality criteria (see Table 7). Evaluations for **direct financial support measures** and for **cluster, technology transfer and networking** measures are more likely to be perceived as being of good quality, while evaluations for softer measures such as **management support** measures or **diffusion** measures are less well assessed in most of the criteria. Other measures, such as mobility show a mixed assessment. The assessment of usefulness of evaluations (see chapter 4 for an in-depth analysis of usefulness) does much less differ between policy measure types, the only exception being diffusion measures for which evaluations as they are performed today are much more likely to be of less usefulness see Table 7.

.

Table 7: Correlation between modality of policy measure and quality and usefulness[30]

	Indirect measures	Direct financial support	Management support	Intermediary bodies	Mobility	Start-ups	Networks, Technology Transfer	Scientific and Industrial Cooperation	Diffusion
Quality									
address TOR	0.1398	0.1532	-0.1714	-0.0341	**0.2207***	-0.1618	0.1041	-0.0119	-0.0797
design appropriate given the objectives	0.0848	**0.2346***	**-0.2555***	-0.065	-0.0466	0.0999	**0.3183***	-0.0403	**-0.1659***
methods satisfy the TOR/purpose	-0.0725	**0.2001***	**-0.1811***	-0.0152	**0.2049***	-0.027	0.0742	**0.2153***	0.0304
application of qualitative methods	-0.01	**0.1704***	-0.1301	0.018	-0.1247	**0.1800***	**0.1862***	0.024	-0.0592
application of the quantitative methods	0.0677	0.0779	0.0092	-0.0857	-0.1057	0.1096	**0.2419***	0.0611	-0.0653
information sources well documented	-0.0315	0.054	**-0.2646***	0.0788	**0.2152***	0.1179	0.0912	0.0546	-0.0645
analysis based on given data	-0.0112	0.1434	-0.0188	-0.0046	**-0.1868***	0.0948	**0.2283***	0.1169	-0.114
cover broader context	-0.0381	**0.1617***	**-0.1922***	0.1282	**0.1730***	0.0017	**0.3056***	0.0222	-0.1252
conclusions based on analysis	-0.0255	0.1441	-0.0795	0.0421	**-0.2326***	0.121	**0.3141***	-0.0345	-0.1291

---

[30] Pearson correlation coefficient; otherwise Spearman correlation coefficient; * significant at 10% level., significant relationships indicated with darker cell shading

	Indirect measures	Direct financial support	Management support	Intermediary bodies	Mobility	Start-ups	Networks, Technology Transfer	Scientific and Industrial Cooperation	Diffusion
Usefulness									
design	-0.111	0.1505	**-0.2554***	-0.0099	0.0452	-0.0572	0.1537	**0.2682***	**-0.2975***
Management / implementation	-0.0923	0.1175	-0.0557	-0.068	-0.1761	0.1865	0.0687	0.1256	**-0.2388***
design/management/implementation of future programmes/measures	-0.1668	-0.0012	0.0644	-0.0463	-0.0955	0.1758	0.0375	-0.046	-0.1192
design/management/implementation of contemporaneous programmes/measures	-0.0216	0.1869	-0.1682	0.0254	-0.219	0.0051	**0.5017***	-0.0505	**-0.2812***
broader policy formulation and implementation	-0.1168	0.0424	-0.1111	0.18	-0.0368	0	0.1879	0.0355	**-0.2303***
Discussion									
discussed within government	0.0343	0.0926	-0.1057	**0.1847***	0.0671	0.0262	0.111	0.0646	-0.005
discussed with participants/stakeholders	-0.0222	0.1521	-0.0331	0.0883	-0.0821	-0.0859	**0.2339***	-0.0423	0.0632

3.2 Characterising evaluation practice – determinants of evaluation design

3.2.1 Introduction

In the following section we explore more in-depth some analytical questions and assumptions. These questions have been formulated in the INNO-Appraisal group meetings including a first Steering Committee session in July 2008. This was based on a first descriptive and explorative analysis of the data set and on ex ante assumptions based on evaluation literature and experience by the project team and the steering committee. As in the previous analyses, the full dataset is used except for those questions that relate to quality, consequence and usefulness, which by definition uses the sample of templates filled in by policy makers only.

3.2.2 Timing of evaluations

A first entry point is to characterise evaluations that differ in their timing. The question is:

Do *ex ante* evaluation use different methods, cover different topics and impacts than *ex post* and interim evaluations?

Overall, it appears that ex ante evaluations and ex post evaluations do not differ significantly in many of those issues. Apparently there is a standard approach to evaluations. Ideally this can be interpreted as providing consistency along the policy measure cycle, allowing for an inter-temporal comparison and indicating that the major evaluation questions are reflected throughout the evaluation cycle.

However, some differences are worth noting. As for topics and impacts covered in evaluations, ex ante evaluations cover more often ‘internal consistency’ and – most obviously – less often ‘goal attainment’ and ‘quality of outputs’ than accompanying, interim and ex-post evaluations. Equally, environmental impacts which tend to be more often considered in ex ante than in other evaluations. This is consistent with the very nature and purpose of ex ante evaluation, i.e. to test the logic of an intervention rather than anticipating goal attainment. At the same time, however, it points to a potential smoke screen effect of covering minority, gender and environmental issues. Those issues appear to be part of general design principles of many interventions and as such are part of ex ante evaluation, as part of the precautionary principle, but are much less often followed up in ex post evaluations. We will come back to this issue when discussing how gender and mobility issues are covered differently depending on who is the sponsor of the evaluation.

3.2.3 Purpose of evaluations: formative vs summative

Further, it is often claimed that summative evaluations are markedly different in their approaches than formative ones, the former geared towards ‘hard’ facts and judgement, the latter more geared towards discourse, monitoring and feedback in order to improve programmes and make actors learn. The first question therefore is:

Are summative evaluation more likely for ex post evaluations, while interim tend to be formative and summative, and are there any related differences as to the target audience of the evaluation?

Table 8 shows indeed that ex ante evaluations are clearly formative (which is only logical), while ex post evaluations are mainly summative. However, it is interesting to note that accompanying evaluations tend to be both formative and summative, indicating that aspects of learning are backed up by judgment and underlying data analysis. Further, the different purpose of formative and summative does not means that the evaluations are geared to different target audiences, our test did not reveal any significant bias.

Table 8: Timing of the evaluation and type of evaluation

Timing of the evaluation		Summative	Formative	both	other	Total
ex-ante	N	0	17	4	1	22
	percentage	0	24.64	7.27	12.5	13.17
accompanying	N	3	8	11	2	24
	percentage	8.57	11.59	20	25	14.37
interim	N	11	40	19	0	70
	percentage	31.43	57.97	34.55	0	41.92
ex-post	N	21	3	21	3	48
	percentage	60	4.35	38.18	37.5	28.74
other	N	0	1	0	2	3
	percentage	0	1.45	0	25	1.8
Total	N	35	69	55	8	167
	percentage	100	100	100	100	100

The slightly different purposes of summative and formative evaluation lead to the question about the use of methods and data sources:

Are there systematic differences between summative and formative evaluation as regards topics, data collection methods and data sources?

Looking at the topics covered it appears that formative evaluations cover significantly more often the following topics: ‘policy/ strategy development’, ‘internal’ and ‘external consistency’ as well as ‘programme implementation efficiency’ and significantly less often input and output additionality. This highlights the very function of formative evaluation which is about understanding the overall fit of a programme into its policy context and the internal logic and efficiency of the programme (see Table 9) and less about concrete, tangible additionalities. This is consistent with the methodological approaches that are applied. Summative evaluations make significant more often use of ‘input output analysis’ and ‘counter factual’ and ‘control group approach’. Formative evaluations seem to have no “unique” methods but rely slightly more often on document analysis and descriptive statistics (even if the latter is not statistically significant) (see Table 10). There are no statistical significant differences for data collection methods and data sources, however, the analysis shows that formative evaluations tend to lean towards qualitative (document analysis) and interactive methods, esp. interviews, focus groups, peer reviews (

Table 11). Formative evaluations are significantly (at 5% level) more often (28% of all) a condition of an external sponsor than summative (6% of all).

Table 9: Topics covered by type of evaluation (summative vs. formative)[31]

Topics covered		formative	summative	Total	chi2/p[32]
Minority issues	N	6	2	8	0.312
	percentage	8.82	5.71	7.77	1
Value for Money/ROI/Cost-Benefit Efficiency	N	11	7	18	0.234
	percentage	16.18	20	17.48	1
Gender issues	N	17	3	20	3.986
	percentage	25	8.57	19.42	0.688
Project Implementation Efficiency	N	22	11	33	0.009
	percentage	32.35	31.43	32.04	1
Output Additionality	N	21	16	37	2.208
	percentage	30.88	45.71	35.92	1
Input Additionality	N	19	19	38	6.888
	percentage	27.94	54.29	36.89	0.13
Behavioural Additionality	N	24	14	38	0.22
	percentage	35.29	40	36.89	1
Quality of Outputs	N	26	16	42	0.535
	percentage	38.24	45.71	40.78	1
Coherence/Complementarity	N	42	13	55	5.629
	percentage	61.76	37.14	53.4	0.265
Programme Implementation Efficiency	N	48	15	63	7.481
	percentage	70.59	42.86	61.17	0.094
External Consistency	N	52	17	69	8.133
	percentage	76.47	48.57	66.99	0.065
Internal Consistency	N	55	19	74	8.08
	percentage	80.88	54.29	71.84	0.067
Policy/Strategy Development	N	55	19	74	8.08
	percentage	80.88	54.29	71.84	0.067
Goal Attainment/Effectiveness	N	50	31	81	3.113
	percentage	73.53	88.57	78.64	1
Outputs, Outcomes and Impacts	N	58	32	90	0.788
	percentage	85.29	91.43	87.38	1
Total	N	506	234	740	
	percentage	744.12	668.57	718.45	
Cases	N	68	35	103	

---

[31] 130 valid cases.

[32] Pearson chi2 /Bonferroni adjusted p-values

Table 10: Data analysis methods used by type of evaluation (summative vs. formative)[33]

data analysis method		formative	summative	Total	chi2/p[34]
Before/After Group Comparison Approach	N	3	2	5	0.095
	percentage	4.35	5.71	4.81	1
Network Analysis	N	6	2	8	0.291
	percentage	8.7	5.71	7.69	1
Input/Output Analysis	N	5	10	15	8.556
	percentage	7.25	28.57	14.42	0.038
Econometric Analysis	N	7	10	17	5.766
	percentage	10.14	28.57	16.35	0.18
Cost/Benefit Approach	N	9	8	17	1.635
	percentage	13.04	22.86	16.35	1
Control Group Approach	N	9	13	22	8.086
	percentage	13.04	37.14	21.15	0.049
Counter-Factual Approach	N	9	14	23	9.796
	percentage	13.04	40	22.12	0.019
Case Study Analysis	N	23	13	36	0.149
	percentage	33.33	37.14	34.62	1
Document Analysis	N	31	13	44	0.577
	percentage	44.93	37.14	42.31	1
Context Analysis	N	41	21	62	0.003
	percentage	59.42	60	59.62	1
Descriptive Statistics	N	53	24	77	0.82
	percentage	76.81	68.57	74.04	1
Total	N	196	130	326	
	percentage	284.06	371.43	313.46	
Cases	N	69	35	104	

---

[33] 104 valid cases.

[34] Pearson chi2 /Bonferroni adjusted p-values

Table 11: Data collection method and sources by type of evaluation (summative vs. formative)[35]

data collection method and sources		formative	summative	Total	chi2/p[36]
Technometrics/Bibliometrics Search	N	0	0	0	.
	percentage	0	0	0	.
Peer Reviews	N	12	3	15	1.464
	percentage	17.39	8.57	14.42	1
Non-participant Surveys	N	15	8	23	0.017
	percentage	21.74	22.86	22.12	1
Focus Groups/Workshops/Meetings	N	37	11	48	4.603
	percentage	53.62	31.43	46.15	0.255
Document Search	N	43	14	57	4.67
	percentage	62.32	40	54.81	0.246
Participant Surveys	N	44	21	65	0.141
	percentage	63.77	60	62.5	1
Existing Surveys/Databases	N	40	27	67	3.724
	percentage	57.97	77.14	64.42	0.429
Monitoring Data	N	51	24	75	0.329
	percentage	73.91	68.57	72.12	1
Interviews	N	56	23	79	3.034
	percentage	81.16	65.71	75.96	0.652
Total	N	298	131	429	
	percentage	431.88	374.29	412.5	
Cases	N	69	35	104	

3.2.4 The meaning of co-sponsors

As seen in section 3.1, innovation policy measures are sometimes co-sponsored by an external party. Within the EU, this is most often the EU Commission through structural funds, and an in-depth case study on the structural fund evaluation practice will shed light on this (see chapter 7). However, more generally, it is important to understand:

Does it make a difference if the policy measure is (co-) financed by an external sponsor in terms of topics and impacts covered and methods used in evaluations?

One starting assumption is that evaluation of external sponsors might tend to differ in the coverage of specific topics and impacts. Indeed, the evidence in the database suggests that external sponsorship makes some difference. Most of all, evaluation of measures with an external sponsor tends to cover more often soft topics and impacts such as minority and gender issues (Table 12) as well as environmental and social impacts more broadly (Table 13). In chapter 7 we will see that this greater concern with softer issues is closely linked to structural fund measures. The tendency to cover those softer issues, in return, is accompanied by a slightly less broad coverage of effectiveness and outcome/ impacts (Table 13).

---

[35] 104 valid cases.

[36] Pearson chi2 /Bonferroni adjusted p-values

Table 12: Topics covered and externals sponsorship (multiple responses)[37]

Topics covered in the evaluation		no external sponsor	external sponsor	Total	chi2/ p[38]
Minority issues	N	2	7	9	16.626
	percentage	1.83	21.88	6.38	0.001
Gender issues	N	12	16	28	23.631
	percentage	11.01	50	19.86	0
Value for Money/ROI/Cost-Benefit Efficiency	N	26	9	35	0.242
	percentage	23.85	28.13	24.82	1
Project Implementation Efficiency	N	45	10	55	1.047
	percentage	41.28	31.25	39.01	1
Input Additionality	N	51	9	60	3.525
	percentage	46.79	28.13	42.55	0.907
Output Additionality	N	51	10	61	2.433
	percentage	46.79	31.25	43.26	1
	N				
Behavioural Additionality	percentage	56	8	64	6.943
	N	51.38	25	45.39	0.126
Quality of Outputs	percentage	61	12	73	3.377
	N	55.96	37.5	51.77	0.991
Coherence/Complementarity	percentage	61	20	81	0.432
	N	55.96	62.5	57.45	1
Programme Implementation Efficiency	percentage	76	24	100	0.334
	N	69.72	75	70.92	1
Policy/Strategy Development	percentage	81	20	101	1.698
	N	74.31	62.5	71.63	1
External Consistency	percentage	79	24	103	0.08
	N	72.48	75	73.05	1
Internal Consistency	percentage	81	25	106	0.193
	N	74.31	78.13	75.18	1
Goal Attainment/Effectiveness	percentage	100	17	117	26.12
	N	91.74	53.13	82.98	0
Outputs, Outcomes and Impacts	percentage	104	22	126	18.499
	N	95.41	68.75	89.36	0
Total	percentage	886	233	1119	
	N	812.84	728.13	793.62	
Cases	percentage	109	32	141	

---

[37] 141 valid cases.

[38] Pearson chi2 /Bonferroni adjusted p-values

Table 13: Impacts covered (aggregated) and externals sponsorship of the evaluated measure[39]

impact		no external sponsor	external sponsor	Total	chi2/p[40]
Environmental	N	19	13	32	7.585
	percentage	17.43	40.63	22.7	0.029
Scientific	N	40	8	48	1.507
	percentage	36.7	25	34.04	1
Social	N	39	17	56	3.108
	percentage	35.78	53.13	39.72	0.389
Technological	N	58	15	73	0.398
	percentage	53.21	46.88	51.77	1
Economic	N	78	24	102	0.146
	percentage	71.56	75	72.34	1
Total	N	234	77	311	
	percentage	214.68	240.63	220.57	
Cases	N	109	32	141	

External sponsors, however, do not impose the use of specific methods or data collection methods on evaluations. The statistical analysis shows few significant differences between evaluations of such measures and the rest as regards methods and data sources, there appears to be a slight tendency for evaluations without an external sponsor to rely slightly more on quantitative methods.[41]

3.2.5 The link of topics and methods

As seen in the previous section in this chapter, evaluations cover a wide variety of topics a set of different categories of impact. To understand the nature of this coverage better, we can see what topics relate more often to what kinds of impact:

Do evaluations that are more interested in certain types of impact concentrate on certain evaluation topics?

To answer this question a set of statistical tests have been conducted. In essence, evaluations that look specifically at scientific and technological impact are very closely linked with a whole range of topics. In other words, those evaluations look very broadly at a whole range of topics (apart from programme efficiency and gender and minority issues). Evaluations with a focus on economic impact have a similar link to a broad variety of topics, but are less concerned with consistency issues (Table 14). This indicates that evaluations focused on scientific, technological and economic impact are complex and broad in their approach. In contrast, those evaluations that are concerned with social and environmental issues are not correlated with most of the evaluations topics, but strongly linked to overall policy and strategy, gender and minority issues (Table 14). This again confirms an earlier finding: there is a distinct set of evaluations that is focused on soft aspects of evaluations, indeed we see a clustering of evaluations between those that cover the soft aspects and those that do not.

---

[39] 141 valid cases, multiple response set

[40] Pearson chi2 /Bonferroni adjusted p-values

[41] In fact the only statistically significant difference is that evaluations with an external sponsor use more often before/after group comparison, but the overall number of cases using this method is rather low (13). Evaluations imposed by an external sponsor use significantly less often participant surveys and interviews.

The analysis presented in (Table 14) also indicates that programme implementation efficiency shows no or even a negative correlation with all three core impact categories, indicating again a clustering of a different sort: evaluations concerned with hard impacts do not cover strongly efficiency issues, effectiveness and efficiency appears to be de-coupled to some extent. [42]

Table 14. Share of evaluations covering a specific combination of topics and impacts[43]

Topics	n	Impacts (binary)				
		Scientific	Technological	Economic	Social	Environmental
External Consistency	163	**36*****	**49****	59	35	**22****
Internal Consistency	164	**38*****	47	60	36	22
Coherence /Complementarity	154	**30****	**38***	44	29	**20****
Goal Attainment/Effectiveness	158	**38***	**53*****	**69*****	37	22
Outputs, Outcomes and Impacts	165	**39*****	**55*****	**70*****	39	23
Quality of Outputs	150	**31*****	**40*****	**50*****	**27***	**17***
Value for Money/ ROI /Cost-Benefit Efficiency	143	**13***	**19*****	**27*****	**15***	9
Programme Implementation Efficiency	155	30	39	**52****	32	19
Project Implementation Efficiency	151	**25*****	**34*****	**42*****	23	**17*****
Input Additionality	148	**28*****	**38*****	**46*****	21	**16***
Output Additionality	150	**29*****	**42*****	**45*****	23	**18*****
Behavioural Additionality	147	**23****	**34*****	40	19	11
Policy/Strategy Development	164	**37*****	**47****	57	**36****	**24*****
Gender issues	151	9	15	19	**15*****	**11*****
Minority issues	144	2	4	6	**6****	**5*****

3.2.6 The link of topics with methods and data collection approaches

The major challenge of evaluations is to apply the appropriate analytical methods and data collection approaches in relation to the topics the evaluator and policy maker are interested in. The question we have tested for our dataset therefore is:

Is there a systematic link of topics covered in evaluations and the methods and data collection approaches used?

Table 15 and Table 16 below display those relations. They show how many evaluations (% of valid pairwise total) use a specific combination of topics and methods/data collection approaches. Consequently, this shows how frequent individual combinations are. The top 10% of combinations are highlighted in grey. For example, in Table 15, 36% of evaluations cover external consistency and at the same time use case study approach. Second, the statistical significance (indicated by the stars) shows if the two characteristics (e.g. topic ‘external consistency’ and method ‘case study approach’) are linked. This is done by comparing two groups (e.g. evaluations that cover ‘external consistency’ and those that do not) in respect to certain characteristics (e.g. using ‘case study analysis’). If Fishers exact chi-square test shows a significant difference in the distribution between both groups (which is the case for topic ‘external consistency’ and method ‘case study approach’), the characteristic (e.g. case studies) is significantly differently often used in both groups of evaluations (those with and without coverage of ‘external consistency’)[44].

---

[42] The statistical tests have also been carried out for a subset of evaluations that is not ex ante, in order to check if the clustering of evaluations into hard and soft or effectiveness vs efficiency is due to a different coverage in ex ante evaluations. However, the differences are negligible

[43] Bold type indicates statistical significant difference at * 10% level, ** 5% level, *** 1% level, based on Fishers' exact chi-square test

The analysis of the combinations reveals some highly interesting results, most of which are intuitive, they correspond to what one would expect in terms of the ways topics are analysed and underpinning data is collected. There are again different groups of topics that are more likely to be analysed with a specific combination of methods and data collection approaches:

- A first group of evaluations is concerned with policy and strategy development issues and looks at external and internal consistency as well as coherence/ complementarity. Those evaluations, by and large, use significantly more often context, document and network analysis as well as before/after group comparison. Consequently, they are fairly strongly correlated with document search, focus groups and interviews. To understand the nature and fit of an intervention, qualitative approaches are essential. In addition, policy development evaluations are also linked to cost/benefit analysis, indicating that the strategic decisions need some quantitative backing.
- The analysis of overall goals (effectiveness and output, outcome and impacts) very generally display a significant correlation with case studies, input/output and descriptive statistics. The data collection methods correlating with this cluster of topics are existing databases (sign.), monitoring data, interviews and participant surveys (sign). The overall effectiveness thus relies on a mix of existing material and rather simple methods to be applied as a standard approach. In addition, general assessments as for outputs and impacts as well as the assessment of the quality of outputs also rely more on peer review, as for holistic judgements as well as content assessment expert knowledge appears to be of key importance.
- For the more complex concept of additionality, the methods and data collection approaches are slightly different. There is some co-linearity between the three types of additionality (input, output or behavioural) in terms of methods and data collection. All three additonalities are correlated significantly with econometric and network analysis as well as counter factual approach. Moreover, for input and output additionality input/ output analysis, before/after group comparison, control group and cost/benefit approach are significant. This is an interesting

---

[44] Since both variables are binary, the argumentation can run both ways. In many cases the direction of the significant difference can be understood from the share of evaluation applying/ covering both aspects. However, since this is not necessarily clear from the share of a combination at the total sample (e.g. evaluations that cover ‘programme implementation efficiency’ and use ‘descriptive statistics’ account for 56% of all evaluations, but only 75% of all evaluations covering ‘programme implementation efficiency’ also use ‘descriptive statistics’, whereas 89% of those evaluations that do not cover ‘programme implementation efficiency’ apply ‘descriptive statistics’In most cases the significance levels of correlation and cross table analysis correspond to each other, in a few cases the significance of the relation is slightly different. E.g. the example of ‘programme implementation efficiency’ and ‘descriptive statistics’ shows a p-value of 0.102 (not significant at the 10% level) for Fisher’s exact test and of 0.0701 (significant at the 10% level) for the Pearson correlation coefficient.

finding that will further be discussed in our case analysis (see Chapter 6 of this report). The data collection methods used for additionality topics are mainly surveys (sign, either non-participants, participant and/or existing ones), monitoring data, interviews and document search.

- A further group of evaluations deals with efficiency issues. Both for programme and project efficiency, case studies and context analysis are important, linked with document search and focus groups, workshops, as it is essential to understand the management structures, processes and practices. Efficiency at the project level, quite logically, is also linked with more sophisticated methods (such as input/output analysis, cost/benefit approaches, network analysis and econometric analysis) that appear to draw on participant survey data.

Table 15. Share of evaluations covering a specific combination of topics and methods (pairwise)[45]

	Case Study Analysis		**Network Analysis**		**Econometric Analysis**		**Descriptive Statistics**		**Input / Output Analysis**		**Document Analysis**		**Context Analysis**		**Before / After Group Comparison Approach**		**Control Group Approach**		**Counter-Factual Approach**		**Cost/Benefit Approach**	
	%	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**
External Consistency	36*	160	16**	146	18	152	61	153	21	154	45***	155	58***	154	11**	151	13	143	14	140	20	153
Internal Consistency	37**	162	17***	149	16	153	60	154	22	155	46***	157	58***	156	10**	153	14	146	13*	141	19	155
Coherence/Complementarity	26	152	15***	140	14	148	47	150	18*	148	42***	149	52***	150	10***	147	11	141	10	136	15	148
Goal Attainment/ Effectiveness	41***	156	17*	144	21	149	71***	150	26***	152	45	152	61	152	9	149	18	142	20	140	21	151
Outputs, Outcomes and Impacts	41***	160	16	146	22**	154	74***	157	26**	156	47	157	63*	156	10	153	18	146	21**	142	23	155
Quality of Outputs	28**	148	14***	140	16**	144	45	143	19***	144	28	145	43***	146	8**	144	12	138	11	135	17**	144
Value for Money/ROI/ Cost-Benefit Efficiency	13	142	6	135	9***	138	25***	139	15***	141	15	139	15	139	7***	138	5	137	7	134	15***	139
Programme Implementation Efficiency	36*	154	15	141	16	144	56	145	20	148	41	151	52*	147	8	144	12**	138	13*	134	16	147
Project Implementation Efficiency	27***	150	13***	141	14**	143	36	143	18***	145	25	144	36***	144	6**	142	7	136	9	132	16***	145
Input Additionality	19	146	9*	138	20***	142	40	139	21***	144	22	139	34	143	8**	139	16***	134	20***	131	18***	141

---

[45] statistical significant difference at * 10% level, ** 5% level, *** 1% level, based on Fishers' exact chi-square test

	Case Study Analysis		**Network Analysis**		**Econometric Analysis**		**Descriptive Statistics**		**Input / Output Analysis**		**Document Analysis**		**Context Analysis**		**Before / After Group Comparison Approach**		**Control Group Approach**		**Counter-Factual Approach**		**Cost/Benefit Approach**	
	%	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**
Output Additionality	21	149	12***	139	19***	144	42*	142	22***	147	23	143	36**	145	8***	142	12**	134	15***	131	19***	144
Behavioural Additionality	23	144	12***	132	16**	138	46***	138	13	140	25	143	28	138	7	135	12	131	18***	128	12	137
Policy/Strategy Development	35**	162	15	148	15	152	56	154	22	156	41*	156	54**	155	9	152	10***	145	12*	141	20	155
Gender issues	11	151	4	140	6	142	16	141	9**	147	14	145	18**	146	6***	142	4	136	2	131	11***	144
Minority issues	1	144	0	135	1	139	6	140	3*	144	5	140	6	143	3**	139	1	136	0	131	4***	141

Table 16. Share of evaluations covering a specific combination of topics and data collection approaches (pairwise) [46]

	Existing Surveys/Databases		Participant Surveys		Non-participant Surveys		Interviews		Focus Groups/ Workshops/ Meetings		Peer Reviews		Technometrics/ Bibliometrics Search		Document Search		Monitoring Data	
	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%	N
External Consistency	59**	157	54	156	20	139	63**	163	45***	155	18***	142	2	148	55***	156	63	147
Internal Consistency	58	158	53	157	19	141	64*	164	44***	157	17	143	2	149	55**	157	61	148
Coherence /Complementarity	47	151	41	150	14	137	50*	154	37***	148	13	138	2	143	49***	148	46	140
Goal Attainment/Effectiveness	64*	154	64***	154	23	137	70	158	42	151	19	139	2	145	55	150	71	142
Outputs, Outcomes and Impacts	66	161	63***	158	23	141	71	163	45	155	19*	143	2	149	57	155	73	147
Quality of Outputs	41	147	42**	146	16	134	46	150	29	146	16***	135	1	142	36	145	45	135
Value for Money/ROI/Cost-Benefit Efficiency	21	141	23**	140	5	132	22	143	14	140	4	134	2**	137	20	140	23**	132
Programme Implementation Efficiency	52	149	49	148	15	136	62	155	44***	149	14	138	2	141	53**	148	60	142
Project Implementation Efficiency	33	146	34**	145	11	131	35	151	29***	147	13***	132	2*	139	34**	145	37	140
Input Additionality	40***	144	34	144	15**	131	35	147	21	141	7	129	1	137	32	142	38	138
Output Additionality	40***	146	35*	146	15**	130	37	150	25	144	8	131	2	139	32	146	37	142
Behavioural Additionality	36	141	40***	140	17***	127	45	146	27	140	9	130	2	132	33	141	38	134

---

[46] statistical significant difference at * 10% level, ** 5% level, *** 1% level, based on Fishers' exact chi-square test

	Existing Surveys/Data bases		Participant Surveys		Non-participant Surveys		Interviews		Focus Groups/ Workshops/ Meetings		Peer Reviews		Technometrics/ Bibliometrics Search		Document Search		Monitoring Data	
	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%	N
Policy/Strategy Development	53	158	48	157	17	141	60*	164	41**	157	15	143	2	149	53***	156	64***	149
Gender issues	19**	145	11	145	5	132	15**	151	14	146	4	133	1	139	18**	145	18	141
Minority issues	6	142	3	142	2	132	3**	144	2	139	1	131	1	136	6*	140	6	136

3.3 Consequences and quality of evaluations

3.3.1 Determinants and consequences of evaluation quality

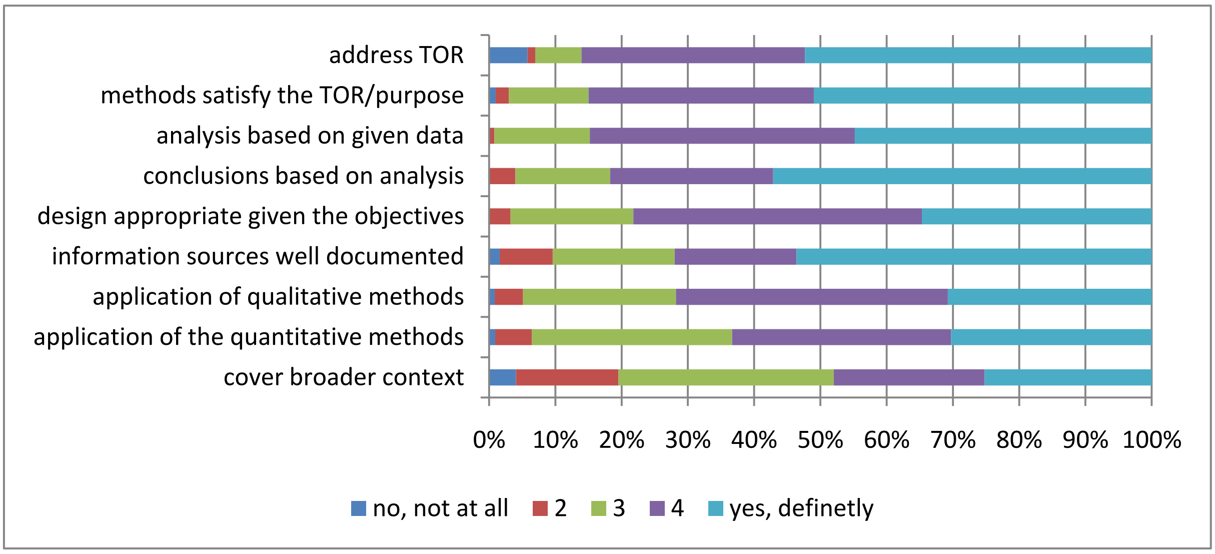
One key purpose of this study is to understand what determines the quality and the usefulness of evaluations. After all, evaluations need to be of use for those commissioning them. The various discussion and steering committee meetings have clearly shown that usefulness of evaluations is the single most important dimension for policy makers. For this reason, there is a separate case study on usefulness of evaluations (chapter 4). This section thus focuses mainly on the quality aspect for and consequences of evaluations .

Quality is not entirely objective, it is also the perception of quality for the policy purpose that is important. Thus, the quality criteria have been assessed by policy makers (i.e. sample B is used here). Quality in this study is defined through a set of criteria. This section depicts the assessment of the evaluations along those individual quality variables. However, given the complexity and high number of variables, it then constructs one composite quality variable in order to better enable statistical analysis around quality and link quality to other aspects of evaluations. The first question thus is:

How do policy makers assess the quality of evaluations?

Our characterisation template included a range of quality indicators. Exhibit 35 shows the distribution for each quality indicator. These present the (perceived) quality of the evaluations based on a five-point Likert scale. The highest satisfaction can be observed for addressing the terms of reference and for the way in which methods satisfy the Terms of Reference. Overall, policy makers seem to be less content with the coverage of the broader context, the application of quantitative and qualitative methods and the documentation of information sources.

Exhibit 35: Distribution across quality categories



Note: sorted in descending order for the top two categories

To better understand quality – or better, the perception of quality – it is important to analyse how quality criteria are linked to other important dimensions of evaluations:

How does quality relate to evaluation characteristics?

Not surprisingly, quality variables are highly correlated with each other, indicating that those evaluations that perform with high quality do so across the board of variables. Most evaluations, thus, are either very good or poor throughout; we cannot see a specific cluster of quality variables. The highest and very plausible correlation can be seen between the use and documentation of data and the way in which evaluation analysis is based on this data. In addition, high quality in the application of quantitative methods is linked to a good and clear documentation of the underlying data. data application of quantitative methods with analysis based on given data and appropriate design with analysis based on given data Table 17.

Table 17: Correlation Matrix quality indicators (pairwise) [47]

		address TOR	design appropriate given the objectives	methods satisfy the TOR/purpose	application of qualitative methods	application of the quantitative methods	information sources well documented	analysis based on given data	cover broader context
design appropriate given the objectives	r	0.2868***							
	n	85							
methods satisfy the TOR/purpose	r	0.326***	0.4523***						
	n	79	100						
application of qualitative methods	r	0.1921*	0.4745***	0.4831***					
	n	79	116	95					
application of the quantitative methods	r	0.3453***	0.4612***	0.4974***	0.333***				
	n	73	108	86	105				
information sources well documented	r	0.305***	0.4014***	0.3475***	0.3009***	0.2069**			
	n	85	123	99	117	108			
analysis based on given data	r	0.3768***	0.5054***	0.4089***	0.4624***	0.5559***	0.307***		
	n	84	121	98	116	108	123		
cover broader context	r	0.1714	0.46***	0.3263***	0.3084***	0.2783***	0.3811***	0.3183***	
	n	82	119	95	113	107	120	120	
conclusions based on analysis	r	0.2729**	0.5597***	0.4317***	0.4884***	0.4206***	0.3119***	0.5763***	0.3413***
	n	85	122	98	115	109	123	124	122

---

[47] Statistical significant difference at * 10%, **5% and ***1 % level based on Spearman's correlation. coefficient

For a further analysis of determinants of quality we take advantage of the composite quality variable that was created and grouped all reports into either high quality or low quality as outlined in the precious section.[48] In total 81 out of 132 evaluations (61%) meet these conditions and thus are regarded as high quality, the other 39% are of low to medium quality. In cases in which the link of individual quality variables and other criteria ate telling, we present the differentiated analysis in addition.

First, our data indicates that quality – the composite variable as just defined – does not systematically differ between evaluations that are conducted by external evaluators and those that are performed internally.[49] Equally, evaluation budgets and the fact that evaluations are build into the design of policy measures do not significantly increase the likelihood of evaluations being of high quality.

Interestingly, evaluations that are done for policy measures sponsored by external or international (co-)sponsors (Table 18) as well as those commissioned by other governmental bodies (Table 19) are apparently of lower quality – or perceived to be of lower quality. The interpretation is not straightforward, evaluations may more likely to be perceived as imposed as conditions of the external sponsorship and thus rated worse by the participating policy makers. Equally, they may be a matter of general routine linked to the external sponsor that does not fit the needs of the specific context. Whatever the reason is, there appears to be room for improvement in the design and conduct of evaluations of co-sponsored measures, a point reconfirmed in our case study on structural fund evaluations (see chapter 7).

Table 18: Evaluation condition of an external/ international (co)sponsor and evaluation quality[50]

		low/ medium quality	**high quality**	**Total**
No	n	29	54	83
	%	56.86	78.26	69.17
Yes	n	22	15	37
	%	43.14	21.74	30.83
Total	n	51	69	120
	%	100	100	100

---

[48] As a reminder: Those evaluations were considered of ‘high quality evaluations’ that score high (>3 on the Likert scale (see Table 17 above) on *all of the following major quality variables* of an evaluation; namely: appropriate design (d3a), analysis clearly based on given data (d8a) and conclusions based on analysis (e1a).

[49] Our methodology, to ask policy makers for their assessment of quality, may have produced a bias here towards internal evaluations, this result, thus, is to be interpreted with care.

[50] Pearson chi2(1) = 6.2961 Pr = 0.012

Table 19: Sponsor of the evaluation and evaluation quality [51]

		low/ medium quality	high quality	Total	chi2/p*
Programme Owner/Manager	n	42	71	113	1.542
	%	82.35	89.87	86.92	0.643
Other Government Department	n	14	13	27	2.277
	%	27.45	16.46	20.77	0.394
Other Public Bodies	n	2	5	7	0.353
	%	3.92	6.33	5.38	1
Total	n	58	89	147	
		113.73	112.66	113.08	
Cases	n	51	79	130	

The tender procedure, however, does seem to make a difference. Evaluations with open tenders – and thus with open competition of various potential evaluators – seem – to be more likely to produce high quality evaluations than those commissioned through closed tenders for a pre-defined group of evaluators (albeit it is not statistically significant). Those evaluations without a tender are much more likely to produce low to medium quality (which is statistical significant at the 5% level; even if the overall number is only 19 in this sample for the assessment of quality).

Table 20: Tender procedure and evaluation quality[52]

		low/ medium quality	high quality	Total
internal	n	4	6	10
	%	8.51	8.96	8.77
no tender	n	12	7	19
	%	25.53	10.45	16.67
closed	n	10	13	23
	%	21.28	19.4	20.18
open	n	21	38	59
	%	44.68	56.72	51.75
other	n	0	3	3
	%	0	4.48	2.63
Total	n	47	67	114
	%	100	100	100

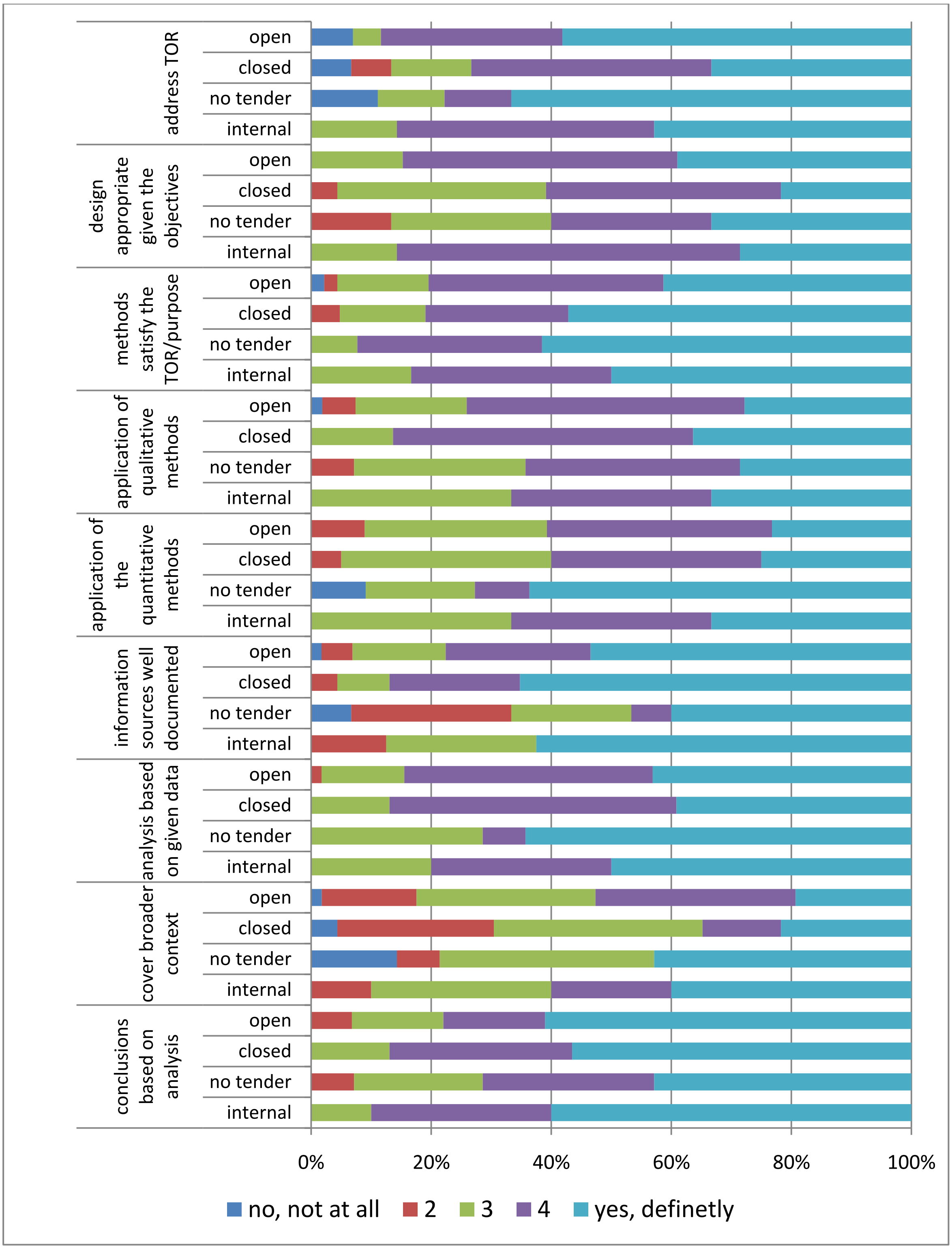
As the tender procedure is an important dimension of evaluations, Exhibit 36 further shows how the various quality aspects vary between evaluations with different tender procedures. Especially the overall design as well as the addressing of the TOR and the coverage of the broader context appear to be more appropriate with evaluations based on open tender.

---

[51] Pearson chi2(1) / Bonferroni adjusted p-values

[52] Pearson chi2(4) = 6.7029 Pr = 0.152

Exhibit 36: Type of tender and evaluation quality criteria



It also appears that ex post evaluations as well as summative evaluations are rated as being of higher quality than ex ante, accompanying and formative evaluations (Table 21 for timing), except in terms of coverage of the broader context. Ex-ante evaluations receive weak rates for the application of qualitative and quantitative methods. This may be linked to the fact that ex post and summative is easier to grasp, while the data basis for formative and ex ante evaluation is less tangible, more qualitative in nature.

Table 21: Median of quality aspects by timing of the evaluation

	ex-ante		accompanying		interim		ex-post		Total	
	Me	n	Me	n	Me	n	Me	n	Me	n
address TOR	4.5	10	4	14	5	39	5	21	5	86
design appropriate given the objectives	4	15	4	17	4	60	5	28	4	124
methods satisfy the TOR/purpose	4.5	10	4	17	4	45	5	24	5	100
application of qualitative methods	3.5	12	4	17	4	58	5	26	4	117
application of the quantitative methods	3	9	4	17	4	56	5	24	4	109
information sources well documented	4	15	4	17	5	59	5	30	5	125
analysis based on given data	4	13	5	19	4	59	5	30	4	125
cover broader context	4	15	4	19	3	60	3.5	26	3	123
conclusions based on analysis	4	14	5	19	4	60	5	30	5	126

There is, in addition, some link between the impacts an evaluation covered and the (perceived) quality of the evaluations. Scientific and technological impacts are more likely to be covered in evaluations that are perceived to be good.

Further, an analysis as to the link of quality on the one hand and methods and data sources used reveals only a few significant results. Evaluations that apply peer review, i.e. rely on the judgment of other experts, appear to most likely perceived as high quality. Similarly, if participant and non-participant surveys are applied evaluations are of higher perceived quality. Expert assessments and the descriptive statistics based on the simple survey method are most credible for policy makers and applied with most clarity. As for methods applied, no clear pattern emerges, some methods (case studies, network analysis, econometric analysis, counter factual approaches and control group approaches) have a higher share of high quality evaluations than other methods, but there is no clear divide between quantitative or qualitative, complex or simple methods. Looking at the various individual aspects of quality rather than the aggregated indicator, we find that case studies, network analysis, econometric analysis and counter-factual analysis are most often positively linked to quality aspects, while input-output analysis, context analysis and document analysis are not systematically linked to quality variables. Apparently the softer topics (context) and methods (document analysis, before/after group comparison) are not often not well enough documented in evaluation studies and the analysis presented is not sufficiently based on the data given.

Quality, of course, is not an end in itself, but we assume that it makes a difference as to the use of the evaluation. Again, for the usefulness of the evaluation we have a separate analysis in chapter 4 of this report. Still, at this stage we can ask the principle question:

Are any differences in the use or target audience and dissemination of evaluations based on its quality?

First, there is some link between the intended internal and external audience for an evaluation report on the one hand and quality on the other hand. Evaluations that are intended for the general public, policy analysts and the target group of the measure itself are more likely to be perceived as high quality (Table 22).

Table 22: Main intended audience and evaluation quality

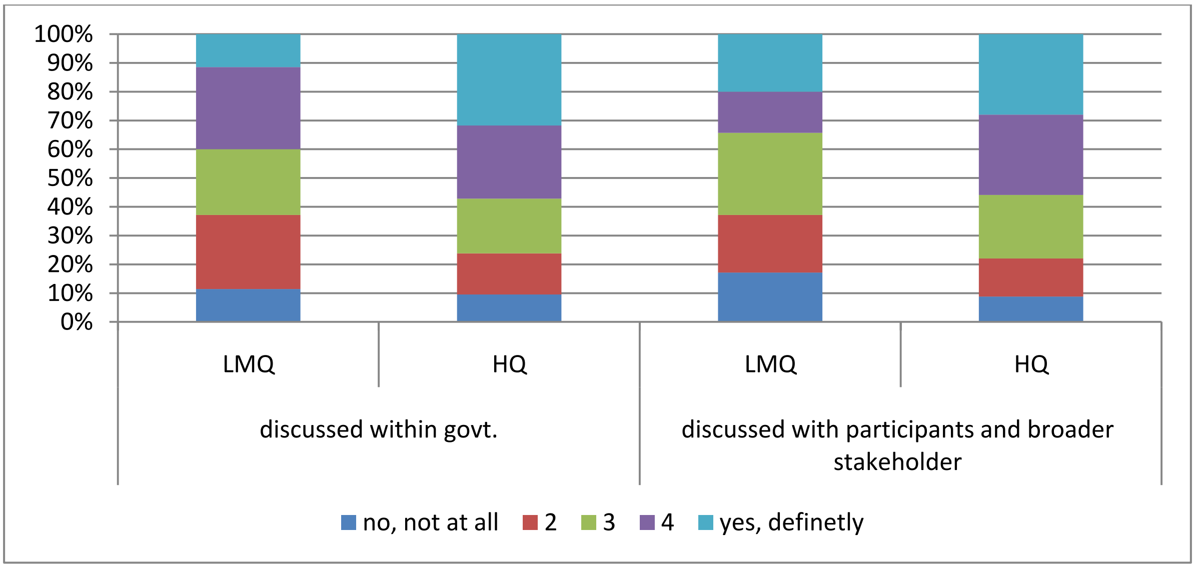
		low/ medium quality	high quality	Total	chi2/p[53]
Policy Makers (Politicians)	n	26	45	71	0.157
	%	52	55.56	54.2	1
Policy Makers (Government Officials)	n	45	73	118	0.001
	%	90	90.12	90.08	1
Programme Management	n	49	77	126	0.727
	%	98	95.06	96.18	1
Auditors/Financial Authorities	n	25	41	66	0.005
	%	50	50.62	50.38	1
Those directly supported by the measure	n	18	40	58	2.244
	%	36	49.38	44.27	1
External/International (co)sponsor of the measure/programme	n	27	27	54	5.45
	%	54	33.33	41.22	0.176
Potential Users of the Measure	n	18	34	52	0.461
	%	36	41.98	39.69	1
Policy Analysts	n	15	48	63	10.602
	%	30	59.26	48.09	0.01
General Public	n	7	28	35	6.68
	%	14	34.57	26.72	0.088
Total	n	230	413	643	
Cases	n	50	81	131	

Second, it also appears that quality has an influence on the likelihood of evaluations to be discussed within government, as high quality evaluations are more likely to be discussed within government and with participants and broader stakeholders than low and medium quality evaluations (Exhibit 37).

---

[53] Pearson chi2(1) / Bonferroni adjusted p-values

Exhibit 37: Distribution of width of discussion and evaluation quality



Looking a bit deeper into the meaning of quality of discussion of an evaluation report, Table 23 differentiates for the different quality aspects. The extent of *discussion within the government* is clearly related to most of the quality variables. Exceptions are the fact if an evaluation addresses the TOR and – at least for wider discussion with participants and stakeholders- if the information sources are well documented and the evaluation covers the broader context. The most influential quality aspect (i.e. highest correlation) is the satisfaction with the ‘application of the quantitative methods’ and ‘conclusions are based on the analysis’. This indicates that clarity of the argument and an easy grasp of the quantitative results as basis for the analysis are key ingredients for the reception of evaluation reports.

Table 23: Correlation coefficients between quality and breadth of discussion (Spearman, pairwise)[54]

		discussed within government	**discussed with participants/stakeholders**
address TOR	r	0.0466	0.1798
	n	71	74
design appropriate given the objectives	r	**0.3248***	**0.2252***
	n	93	98
methods satisfy the TOR/purpose	r	**0.3460***	**0.2319***
	n	84	88
application of qualitative methods	r	**0.2997***	**0.2812***
	n	87	92
application of the quantitative methods	r	**0.3735***	**0.3228***
	n	83	87
information sources well documented	r	**0.1782***	0.05
	n	93	98
analysis based on given data	r	**0.3303***	**0.2435***
	n	91	97
cover broader context	r	**0.2129***	0.1092
	n	90	95
conclusions based on analysis	r	**0.3518***	**0.3662***
	n	93	99

3.3.2 Determinants of evaluation consequences

Evaluations may have all sorts of impact. The may – should – lead to learning more generally and they should support decision making for later stages or new programmes. As seen in Chapter 2 and the previous section of this chapter, we have broken down the broad concept of consequences into five different categories. Exhibit 34 shows that the most frequent consequences are either minor modification of the measure/ programme (n= 61) or its expansion/ prolongations (49). More severe consequences such as termination of the programme/ measure (4), a major re-design (11) or mergers (14) are very seldom. More frequent is the re-design (23) of another measure.

The question that arises from having data on those consequences is obviously:

How do consequences relate to methods and data used in evaluation?

In other words, do we see any systematic characteristics of evaluations that increase or decrease the likelihood for certain consequences to occur? First, we find no systematic correlation between major re-design on the one hand and any kind of topic covered, methods used or data sources. This is partly due to the low number of evaluations that led to major re-design. Further analysis of those cases also seems to indicate (note that the number is too low for hard evidence) that major redesign is often a result of a political process, of changes in directions that are not attributed to evaluation results. Second, expansion/prolongation of measures is linked to a set of topics and methods. It appears that prolonging or extending a measure is based on evaluation that cover consistency, goal attainment and output, outcomes and impacts and, interestingly and strongly correlated, behavioural additionality. The latter indicates that behavioural additionality is important to show effects of measures that are long running and complex, with hard effects taking longer to realise and behavioural changes to be observed as an interim step for further impacts (see case study on BA, chapter 6). Prolongation and expansion are further also correlated with more simple methods such as case study and descriptive statistics based on interviews and surveys, again, it appears that clear and simple messages are more likely to lead to expansion or prolongations.

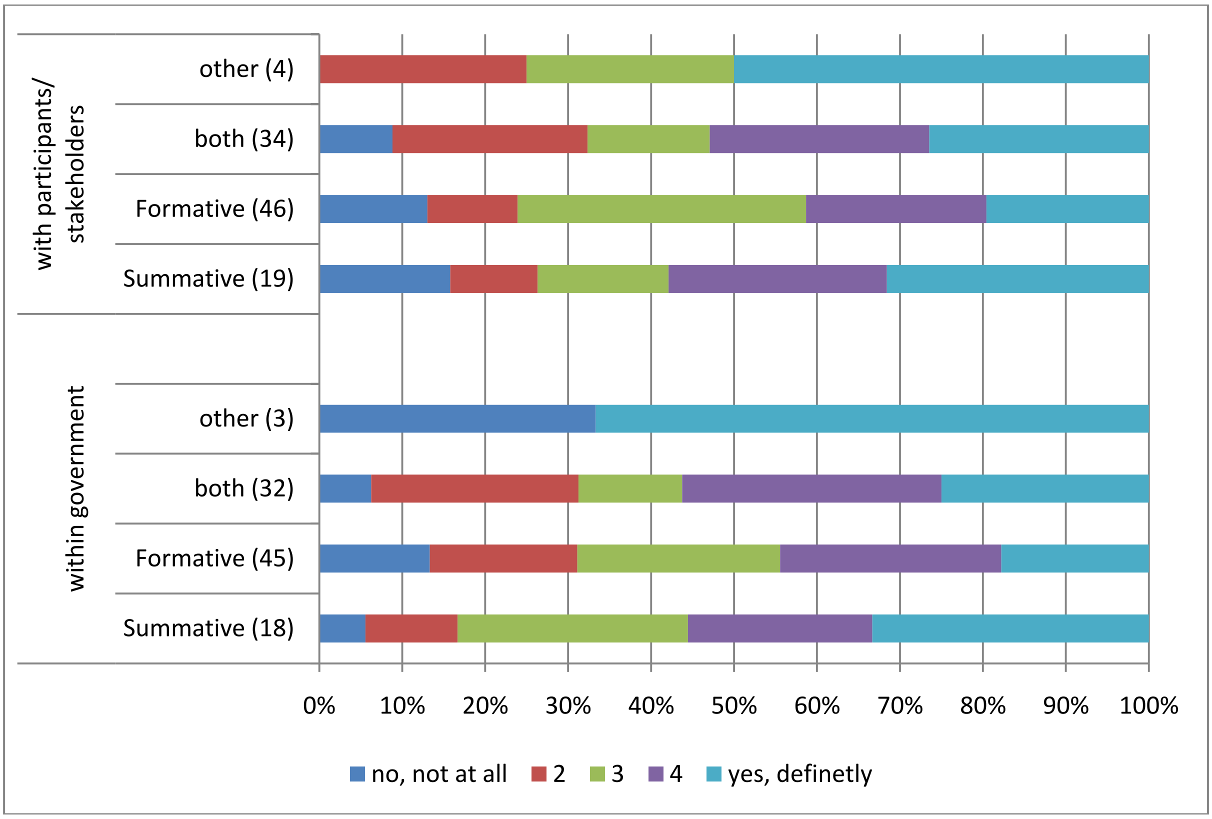
---

[54] * significant at the 10% Level

Are there differences between summative and formative evaluations in terms of discussion about and consequences of evaluations?

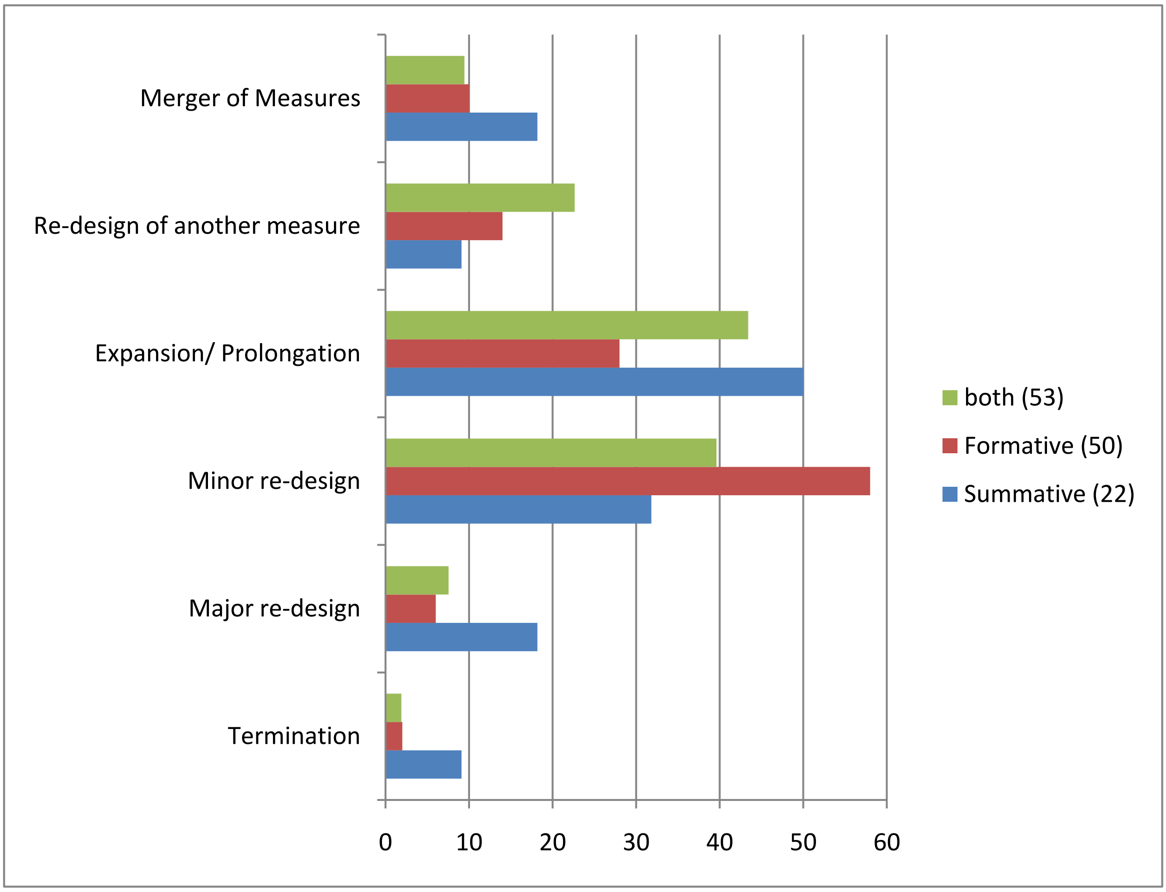
Exhibit 38 shows that evaluations which are (at least partly) summative tend to be more often widely discussed within government and with participants/ stakeholders than formative evaluations (the category other is ignored due to a very low frequency). Even if the differences are not statistically significant, it appears the results of summative evaluations, with clear 'numbers' and messages, are better suited for wider discussion, while the virtue of formative evaluation is not so much their dissemination, but the learning within the process itself.

Exhibit 38: Width of discussion by type of evaluation in % (n in brackets)



Consequences do not differ between those purposes of evaluations in a statistical significant manner. Nevertheless, the general distribution indicates that within the sample summative evaluations tend to have more often caused severe consequences such as termination, major re-design or merger of measures. Also, the expansion is more frequent for this type of evaluation. In contrast, formative evaluations tend to cause more often minor modifications of the measure (Exhibit 39).

Exhibit 39: Consequences by type of evaluation (in %; category other omitted)



Finally, we can come back to the quality aspect discussed in the previous sub-section and link quality of evaluations to the consequence, the idea being that higher quality of evaluations might more often lead to consequences as a result of the evaluation:

Is there a positive relation between the quality of an evaluation and the consequences the evaluation has for the policy makers?

If we look at the composite quality indicator, the only statistical significant difference is that high quality evaluations lead more often to the expansion/ prolongation of a measure (see Table 24).

Table 24: Consequences of high and low/medium quality evaluations

		low/ medium quality	**high quality**	**Total**	**chi2/p**[55]
Termination	n	3	1	4	2.301
	%	5.88	1.23	3.03	0.776
Major re-design	n	3	8	11	0.654
	%	5.88	9.88	8.33	1
Minor re-design	n	23	38	61	0.042
	%	45.1	46.91	46.21	1
Expansion/Prolongation	**n**	**9**	**40**	**49**	**13.504**
	%	**17.65**	**49.38**	**37.12**	**0.001**
Re-design of another measure	n	9	14	23	0.003
	%	17.65	17.28	17.42	1
Merger of measures	n	5	9	14	0.056
	%	9.8	11.11	10.61	1
Total	n	52	110	162	
Cases	n	51	81	132	

To dig slightly, deeper, Table 25 displays the correlation between individual quality indicators and consequence categories (whereby we have to keep in mind the very low frequencies for some of the consequence categories, see above). Evaluations with a higher (perceived) quality have a positive and significant relation with the expansion/ prolongation of a programme/ measure. In this sense 'good' evaluations seem to induce the extension of programmes/ measures, or – vice versa – an evaluation with positive recommendation of an evaluation (which might result into the extension/ prolongation) is more often assumed by policy makers to be of good quality. The most striking feature of Table 25 is that the termination of a measure is significantly negatively correlated to the quality of the analysis in terms of being based on the given data. All this appears to confirm an earlier finding, whereby strong, major decisions of re-design or termination may not so much depend on evaluation results and quality, but based on other consideration such as change of policy priority. However, it is important to point out again that there are very (too) few cases in which an evaluation caused a termination. The most influential quality aspects are satisfactory methods relation to the objectives of the in terms of TOR which are most linked to minor re-design and the satisfactory application of qualitative methods most linked to expansion/ prolongation.[56]

---

[55] Pearson chi2(1) / Bonferroni adjusted p-values

[56] Those relations hold – in principle - true also if we exclude ex ante evaluations.

Table 25: Correlation coefficients between individual quality indicators and consequences of the evaluation (Spearman, pairwise)[57]

	n	**Termination**	**Major re-design**	**Minor re-design**	**Expansion/ Prolongation**	**Re-design of another measure**	**Merger of measures**	**Any consequence**
address TOR	86	-0.0847	-0.1114	0.0902	0.1648	0.1178	**-0.2122***	0.0372
design appropriate given the objectives	124	-0.0486	0.0132	**0.2134***	**0.2382***	0.0805	-0.0505	**0.2725***
methods satisfy the TOR/purpose	100	-0.1441	-0.0653	**0.2081***	0.1555	-0.0754	-0.1005	0.1205
application of qualitative methods	117	-0.135	-0.0058	0.1325	**0.2999***	0.0913	0.0145	**0.2311***
application of the quantitative methods	109	-0.1413	-0.1457	0.1047	**0.2952***	0.0733	-0.1129	0.076
information sources well documented	125	-0.1125	**-0.1621***	0.1446	0.058	0.0841	-0.0562	0.0377
analysis based on given data	125	**-0.2098***	-0.0534	**0.1627***	**0.2884***	0.1396	-0.1141	**0.2194***
cover broader context	123	**-0.1589***	-0.0531	0.1478	**0.2800***	0.0533	-0.1123	**0.1550***
conclusions based on analysis	86	0.0473	0.0824	**0.1794***	**0.2403***	0.0377	-0.0245	**0.2565***

Analysing how evaluations are used, it is interesting to see if there is a link between the scope of discussion about evaluation results on the one hand and the consequences on the other hand:

Does the depth and scope of discussion relate to the consequences?

It is obvious that the database cannot sufficiently answer this question, as many aspects will influence the breadth of discussion as well as the final decision. However, there are stable positive relations between the intensity and scope of the discussion on the one hand and consequences on the other hand for consequences in general and for the two most frequent consequences minor re-design and for extension and prolongation (Table 26). The less frequent consequences (termination, re-design of another measure and major redesign) are not linked to the mode of discussion. Interpretation of the latter is not straightforward, as all three of them show low numbers and thus statistics are problematic.

---

[57] Bold type indicates statistical significant difference at * 10% level based on Spearman's correlation coefficient.

Table 26: Correlation coefficients between discussion indicators and consequences of the evaluation (Spearman, pairwise)[58]

	discussed within government (n= 98)	discussed with participants/ stakeholders (n=103)
Termination	-0.0673	0.0129
Major re-design	0.1075	-0.0362
Minor re-design	**0.2340***	**0.2116***
Expansion/Prolongation	**0.3229***	**0.3454***
Re-design of another measure	-0.1478	-0.0293
Merger of measures	0.1055	-0.001
Any consequence	**0.3683***	**0.2926***

3.4 Cluster of evaluations – an exploration

In a final step of the analysis a hierarchical cluster analysis was conducted, an explorative method which tries to find groups of similar evaluations (clusters) within our sample.[59] The cluster analysis groups evaluations based on information about their characteristics and characteristics of the policy measures they evaluate. In detail the following aspects are included:

- Appraisal characteristics: Timing and purpose, reason (condition of an external/ international (co)sponsor), topics and impacts covered, analysis and data collection method applied, main audience.
- Appraised measure characteristics: modality and target group of measure.

This cluster analysis should not imply two simplified types of evaluation. Rather it should be seen as a further attempt to categorise evaluations beyond those uni-dimensional characteristics normally used for evaluations – as within this study so far.

Two clusters emerge, cluster 1 (38 cases) and cluster 2 (42 cases). Those emerging clusters are not entirely clear cut, but some tendencies are to be observed (see Table 27). It seems that cluster 1 is more populated by ex ante evaluations and is concerned with programme efficiency issues and, by nature, more often based on qualitative methods. Cluster 2 appears to be more ex post and interim, being broader in its coverage and more concerned with different forms of outcome/impact, thereby mobilising more quantitative approaches and oriented towards the policy community rather than the beneficiaries. This cluster of evaluations is more often used for decision about prolongation or re-design of measures.

---

[58] Bold type indicates statistical significant difference at * 10% level based on Spearman's correlation coefficient

[59] Since these clusters are to be aspects of quality, usefulness etc. **sample B** is used.

Table 27: Comparison of two clusters of evaluation

	Cluster 1 (38 cases)	Cluster 2 (42 cases)
Timing	includes the majority of ex-ante and interim evaluations. In absolute numbers it is dominated by interim evaluations	encompasses the majority of accompanying and ex-post evaluations. In absolute numbers it is dominated by ex-post and interim evaluations
Purpose	No clear difference	
Topics	look significantly more often into program implementation efficiency	significantly more often concerned with goal attainment/ effectiveness, output/ outcome/ impact and quality of output as well as the additionality aspects (input, output, behavioural). Cluster 2 has a higher average number of topics covered (8.9) than cluster 1 (6.4).
Impacts covered		look significantly more often into each impact category – this is also reflected by the average number of impact categories considered (0.9 for cluster 1 and 2.4 for cluster 2)
Methods used	significantly more often based on document analysis	use more often econometric analysis, descriptive statistics and counter –factual approach
Main intended audience	more often oriented towards external/international (co)sponsors	are more often targeted at policy analysts
Modality	No important differences	
Target groups	No important differences	
Quality		significantly more often of high quality (using our aggregated quality indicator), significantly more often perceived of being of a high quality in terms of addressing TOR, application of quantitative methods, analysis based on data given and conclusion based on analysis, also more favourable in terms of appropriate design, application of qualitative methods and covering broader context but not significant.
Usefulness	Differences statistically not significant, all in all appear to be a bit more useful for design and management/ implementation of the measure appraised	Appears to be more useful for design and implementation of other, future programmes
Discussion	No significant differences	
Consequences		contribute significantly more often to the expansion/ prolongation and the re-design of another measure
Geographical distribution[60]	Some dominance of this type in AT, CZ,GR, HU, MT, SI, SK,	Some dominance of this type in BE, DE, ES, FI, IT, NL, PL and UK

---

[60] One has to keep in mind the low frequency in some countries and also our caveat that we only refer to the evaluations of measures that are flagged out as such in Trendchart, thus the country representation here does not correspond to the entirety of available evaluations (see section 2 and 3.1 of this report)

References

Jann, B. (2005a), Tabulation of multiple responses, The Stata Journal 5(1), pp. 92-122

Jann, B. (2005b), Einführung in die Statistik, Oldenbourg: München

Corder, G.W., Foreman, D.I. (2009), Nonparametric Statistics for Non-Statisticians: A Step-by-Step Approach Wiley: Hoboken

R. G. Lomax (2007), An introduction to statistical concepts. Lawrence Erlbaum: Mahwah

STATA (2007), STATA Multivariate Statistics Reference Manual, Release 10. Stata Press: College Station.

Řezanková, H. (2009). Cluster analysis and categorical data. *Statistika***,** 216-232.

This page has been intentionally left blank.

Part II Chapter 4

Usefulness of Evaluations

This chapter examines the issue of usefulness (or utility) of evaluations. It notes the role of evaluation (as a key policy tool) within the governance of the overall policy mix in enabling policy makers to judge the effectiveness and efficiency of their interventions. It reviews the literature to determine the possible nature of the idea of usefulness and relates these to the purposes of evaluations. Potential ways in which the utility of evaluations may be enhanced are also discussed.

Two lines of analysis are followed: looking for evidence of utility provided by the survey responses and testing of hypothesised links between utility and other survey database variables. Usefulness is defined through the construction of a proxy indicator. The analysis examines a number of the database variables for links with usefulness.

Overall, the results of the analyses present a mixed picture, confirming some expectations yet failing to confirm or even refuting other expectations. Nevertheless, the results do tend to support the overall conclusion (also based on the direct input of policymakers in the field) that usefulness is a highly subjective and context specific issue.

Paul Cunningham and Abdullah Gök[61]

---

[61] The lead authors for this section of the report are as given above, with contributions from all other team members at different parts of its production, conclusions and interpretations have been approved and are shared by the group.

Table of Contents

Table of Contents 88

List of Tables 89

Executive Summary 91

1 Introduction 94

2 Usefulness and utility: considerations 95

3 How to increase the usefulness of an evaluation 96

4 Potential areas for analysis in the database 98

4.1 Proxies and other indicators of usefulness from the templates 98

4.2 Testing of hypothesised links between utility and other database variables 100

4.2.1 Basic statistics from the analyses 100

4.2.2 Use of external evaluators 101

4.2.3 Timing of evaluation 101

4.2.4 Purpose of evaluations: Summative versus formative 102

4.2.5 Evaluation of Structural Fund measures and portfolio evaluations 102

4.2.6 Conditional evaluations 103

4.2.7 Planned evaluations 103

4.2.8 Dedicated budget for appraisals 104

4.2.9 Topics and utility 104

4.2.10 Impacts examined by the appraisal and usefulness 107

4.2.11 Sponsor effects 109

4.2.12 Availability and language of report 109

4.2.13 Methodological effects on utility 110

4.2.14 Audience effects 114

4.2.15 Breadth of discussion and usefulness 116

4.2.16 Tender procedures and usefulness 116

4.2.17 Consequences of the appraisal and usefulness 117

4.2.18 Country-specific analyses 118

4.2.19 Usefulness and type of innovation support intervention 119

4.2.20 Usefulness and quality 120

5 Conclusion 121

List of Tables

Table 28: Appraisals containing recommendations and their policy target 100
Table 29 Usefulness composite indicators (median) 101
Table 30: Usefulness and use of internal vs external evaluators 101
Table 31: Timing of appraisal and usefulness 102
Table 32: Usefulness of summative and formative appraisals 102
Table 33: Usefulness of Structural Fund and portfolio evaluations 103
Table 34: Usefulness and conditionality 103
Table 35: Planning and usefulness 104
Table 36: Dedicated budget and usefulness 104
Table 37: Topics covered by the appraisal and usefulness (1) 104
Table 38 Topics covered by the appraisal and usefulness (2) 105
Table 39: Topics covered by the appraisal and usefulness (3) 105
Table 40: Topics covered by the appraisal and usefulness (4) 106
Table 41: Topics covered by the appraisal and usefulness (5) 106
Table 42: Topics covered by the appraisal and usefulness (6) 106
Table 43: Topics covered by the appraisal and usefulness (7) 107
Table 44: Types of impact examined and usefulness (1) 108
Table 45: Types of impact examined and usefulness (2) 108
Table 46: Types of impact examined and usefulness (3) 109
Table 47: Appraisal Sponsor and usefulness 109
Table 48: Level of availability of appraisal report and utility (1) 109
Table 49: Language of appraisal report and utility (2) 110
Table 50: Data analysis methodologies and usefulness (1) 110
Table 51: Data analysis methodologies and usefulness (2) 110
Table 52: Data analysis methodologies and usefulness (3) 111
Table 53: Data analysis methodologies and usefulness (4) 111
Table 54: Data analysis methodologies and usefulness (5) 112
Table 55: Data analysis methodologies and usefulness (6) 112
Table 56: Data collection methodologies/sources and usefulness (1) 112
Table 57: Data collection methodologies/sources and usefulness (2) 113
Table 58: Data collection methodologies/sources and usefulness (3) 113
Table 59: Data collection methodologies/sources and usefulness (4) 113
Table 60: Data collection methodologies/sources and usefulness (5) 114
Table 61: Main intended audience for the appraisal and usefulness (1) 114
Table 62: Main intended audience for the appraisal and usefulness (2) 115
Table 63: Main intended audience for the appraisal and usefulness (3) 115
Table 64: Main intended audience for the appraisal and usefulness (4) 115
Table 65: Main intended audience for the appraisal and usefulness (5) 116
Table 66: Breadth of discussion and usefulness 116
Table 67: Tender procedures used and usefulness 117
Table 68: Consequences of appraisal and usefulness (1) 117
Table 69: Consequences of appraisal and usefulness (2) 118
Table 70: Consequences of appraisal and usefulness (3) 118
Table 71: Country specific analysis 119

Table 72: Modality of policy measure and usefulness 120
Table 73: Correlation coefficients between quality and usefulness indicators (Spearman (r), pairwise) 121

Executive Summary

This chapter examines the issue of usefulness (or utility) of evaluations. It begins by setting the broader context of policy interventions within a policy mix and the accompanying need for policy makers to be able to judge the effectiveness and efficiency of their interventions through the use of a range of governance tools, including appraisal, monitoring and evaluation. It is clear, from the policy mix concept, that the information gained from these tools should not be restricted to the subject of the assessment but should also be relevant to the design and operation of contemporaneous or subsequent policy instruments: such requirements define the issue of usefulness.

The discussion then moves to considerations, largely derived from the literature, of what is meant by usefulness and utility in the context of the evaluation of innovation support measures. Three major purposes for evaluation are identified: operational learning, policy feedback and system impact. Overall, it is suggested that, to be useful, evaluations must provide information on:

- the effectiveness of design
- the effectiveness of management
- the effectiveness of implementation
- the effectiveness of the evaluation itself
- the achievement of objectives
- the broader impacts of the instrument.

However, it is recognised that usefulness may also be impacted by other factors such as audiences and sponsor demands.

A number of factors are then examined whereby the utility of evaluations may be increased. Possible routes include increasing the rigour (and hence 'quality') of an evaluation, obtaining the compliance and trust of stakeholders, improving the transparency of methodologies (assuming an informed audience of policy makers is present), and the use of clear and measurable objectives. The incorporation of evaluation into the overall policy cycle is seen as a clear route to improving the usefulness of its outcomes.

The chapter next deals with the approaches employed in the analysis of the survey results to determine the extent of usefulness of the evaluations reviewed. Two lines of analysis were followed: looking for evidence of utility provided by the responses and testing of hypothesised links between utility and other database variables. As the questionnaire did not specifically pose a direct question on the usefulness of the evaluation (which would have prompted highly subjective responses unsuitable for quantitative analysis), it was necessary to develop a proxy for usefulness based on the extent to which the evaluation report's recommendations had been useful (a point addressed by specific questions in the questionnaire template). This proxy indicator (for overall usefulness) could be broken down into internal utility (relating to changes to the programme under appraisal) and external utility (relating to changes to contemporaneous or subsequent programmes).

The analysis then examined a number of the database variables for links with usefulness. The main points to emerge were:

- 84% of evaluations examined had contained recommendations, with an almost equal balance between internal recommendations (relevant to the subject programme) and external recommendations (relevant to future programmes or to broader policy formulation).
- Evaluations addressing internal aspects of the programme had a slightly higher usefulness than those addressing external aspects.
- Significant positive correlations with at least one aspect of usefulness were identified for:
    - The use of an open tendering process when commissioning and evaluation
    - The use of external evaluators
    - The timing of the evaluation (*ex ante*, interim, *ex post*, etc.)
    - Summative over formative evaluations
    - Non-Structural Fund evaluations (i.e. a negative correlation between Structural Fund evaluations and utility)
    - Non-portfolio type evaluations (i.e. a negative correlation between portfolio type evaluations and utility)
    - Non-conditional evaluations (i.e. a negative correlation between conditional evaluations and utility)
    - Evaluations that examined the topics of goal attainment and effectiveness and policy/ strategy development
    - Evaluations that examined scientific impact and technological impact on the participants and beyond
    - Evaluations that employed case study analysis; participant surveys; interviews; focus groups/workshops and meetings; peer review
    - Evaluations that resulted in a minor redesign or expansion/prolongation of the measure
    - Evaluations sponsored by programme managers, other government departments or other public bodies
    - Evaluations **not** conducted primarily for auditors/financial authorities
    - Evaluations whose reports were published in English
    - Certain dimensions relating to the quality of the evaluation
- Negative correlations with at least one aspect of usefulness were observed for:
    - Evaluations that examined input additionality and environmental impacts
    - Evaluations that employed input/output analyses; context analysis; group comparison approaches; cost/benefit approaches; existing surveys and databases
- No significant correlations with any aspect of usefulness were detected for:
    - Evaluations planned during the design of the measure
    - Presence of a dedicated budget for the evaluation
    - Evaluations conducted primarily for policymakers (government officials) and programme management
    - Evaluations that examined outputs, outcomes and impacts; quality of outputs; value for money; programme/project implementation efficiency
    - Evaluations that employed monitoring data
    - Evaluations that had wider levels of availability
    - Evaluations where a major redesign of the measure resulted

- External usefulness was more highly rated in Germany and the Netherlands, whilst internal usefulness was more highly rated in Greece, Sweden and the UK
- The evaluations of measures for science-industry cooperation were significantly more useful across all categories of usefulness. Evaluations of measures aimed at the creation of start-ups and spin-offs were also significantly useful (external and overall).

Whilst a number of the statistically significant associations between usefulness and the survey variables were anticipated, it is harder to explain some of the negative correlations or where no correlations were detected. Several of the latter might be explained by the relatively low number of cases available within the analysis, whilst the prevalence of Structural Fund evaluations within the sample could also provide an explanation.

In conclusion, the results of the analyses present a mixed picture, confirming some expectations yet failing to confirm or even refuting other expectations. As with most research endeavours, it is clear that further investigations are required into the aspect of usefulness and it is hoped that this study offers a valuable starting point. Nevertheless, the results do tend to support the overall conclusion (which is also based on the direct input of policymakers in the field) that usefulness is a highly subjective and context specific issue and that, as a broad rule of thumb, an evaluation may be considered useful if it delivers the Terms of Reference in a consistent manner and if it provides actionable recommendations and delivers value for money. Usefulness can be defined as the degree to which there is feedback on policy and if the evaluation process delivered some degree of policy learning.

1 Introduction

A recent development in studies of R&D and innovation policy is the increasing attention paid to the concept of policy mixes[62]. Policy mixes are defined as the "combination of **policy instruments**, which **interact** to **influence the quantity and quality of R&D investments** in public and private sectors". Policy instruments encompass "All programmes, organisations, rules and regulations with an active involvement of the public sector, which intentionally or unintentionally affect R&D investments. This usually involves some public funding, but not always as, for example, regulatory changes affect R&D investments without the intervention of public funds". The issue of interaction between policy instruments is also core to the policy mix concept and relies on the fact that "the influence of one policy instrument is modified by the co-existence of other policy instruments in the policy mix"[63].

One of the major conclusions of the policy mix study conducted for DG Research[64] was that the process of governance plays a central role in setting the overall shape and in defining the development of the policy mix. With particular relevance to the INNO-Appraisal study, the governance process pays particular attention to the effectiveness of the innovation policy mix. It does this notably through the processes of evaluation, monitoring and *ex ante* review, together with other related policy-making tools, applied at the level of the policy mix, to significant sectoral elements of it, or to individual component parts (instruments) that operate within it. Feedback from all these levels thus informs policy and directs the future composition of the innovation policy mix

As a consequence, the policy mix is clearly a product of an evolutionary development process within which governance is a major driver and, in turn, within which the processes of review (of the relevant innovation system or its component elements) and assessment are important selection factors, while the balance between various instruments is a product of both the impacts of existing measures, available resources and wider policy needs. As systems are dynamic entities, effective policy mixes must also adapt and change and governance (and, hence, the tools it employs) must be responsive to such changes. Thus, the use of governance processes results in a policy mix that is a dynamic entity, constantly in flux and changing to meet the needs of the innovation system.

The INNO-Appraisal Project has focused on a number of specific governance tools, namely those relating to evaluation. These encompass *ex ante* assessment, monitoring and interim evaluation and *ex post* evaluation. A major requirement for these tools, indeed one of their primary *raison d'etres*, is that they should be able to provide information to policy makers on the performance (either potential – in the case of *ex ante* assessment, or actual – in the other cases) of the policy instruments to which they are applied. Moreover, the information gained from such assessments of policy instruments may not be restricted to the subject of the assessment but also be of use to the design and operation of contemporaneous or subsequent policy instruments. In short, appraisals and their outcomes must be useful.

---

[62] See, for example, Nauwelaers, C, Boekholt, P, Mostert, B, Cunningham, P, Guy, K, Hofer, R, and Rammer, C, "Policy Mixes for R&D in Europe", Report commissioned by European Commission, Directorate-General for Research, UNU-MERIT, Maastricht. May 2009.

[63] Op. cit.

[64] Cunningham, P.N., Monitoring and analysis of policies and public financing instruments conducive to higher levels of R&D investments: The "POLICY MIX" Project: Thematic Report: Governance, European Commission, January 2008.

2 Usefulness and utility: considerations

In the context of an evaluation, what do we mean by usefulness and utility[65]? As evaluation is a governance tool used by policymakers, its utility and usefulness clearly are defined by the purposes for which it is applied. Miles and Cunningham[66] note that, while evaluations are conducted for a range of purposes, according to diverse rationales, they are essentially a learning tool and are carried out “in order to generate lessons for future practice, so that things may be done better in the future”. Drawing on work by VINNOVA, three ways can be identified in which evaluation may assist learning:

- Operational learning: where evaluation is used as a management feedback tool to improve the effectiveness, efficiency and quality of policy intervention by the organisation responsible for the implementation of the programme. Lessons are sought on how organisations (ministries, agencies, etc.) can improve in designing, managing and implementing programmes, while the evaluation process itself may be improved in the future through lessons learned in the present evaluation.
- Policy feedback: The “traditional” idea in which evaluation is used to determine the outcome and impacts of policy measures and programmes. Such evaluations may also be used to check whether, and the extent to which, the objectives of programmes have been achieved. Not only can this type of evaluation contribute to the design of future programmes, it also allows policy makers to test their initial assumptions about the identified bottlenecks and market or system failures which prompted the policy intervention in the first place.
- System impact: By guiding the design and formulation of intervention policies and programmes, evaluations conducted at this level help to improve the efficiency and effectiveness of national innovation systems. Answers are sought to broader level questions concerning the innovation system such as when certain interventions are appropriate, which complementary programmes should be used and when, what is the appropriate policy mix needed to achieve the desired effects, etc.[67]

These approaches portray an increasingly sophisticated rationale for the application of evaluation, tracing the evolution of a culture of evaluation from the restricted management oriented view of programme administration, through evaluation as a justificatory process to a means to obtain a holistic and systemic view of innovation policy intervention. However, it does not follow that evaluations of multiple programmes are required in order to derive information on the performance of the overall policy; depending on the scope and methodologies applied, it could be possible to discern the broader, systemic effects of a single policy instrument. Thus, although INNO-Appraisal was, of necessity, restricted to evaluations of single innovation support instruments (with the exception of ‘portfolio’ evaluations), the data may demonstrate examples of evaluations where methodologies have sought, and captured, information on such systemic effects.

To summarise so far, the main issues that provide usefulness and utility to policy makers concern the information sought by the evaluation, namely:

---

[65] Here, the terms ‘usefulness’ and ‘utility’ are used synonymously. Utility is not used in its economic sense, but is taken to mean the fact, quality or character of being useful or fit for a purpose, while usefulness implies being helpful in effecting a purpose.

[66] Miles, I. and Cunningham, P.N., *Smart Innovation: A Practical Guide to Evaluating Innovation Programmes*, Supporting the monitoring and evaluation of innovation programmes: A step towards European inno-policy governance”, Report to DG Enterprise and Industry, January 2006.

[67] Miles and Cunningham 2006, adapted from VINNOVA, 2004.

- Information on the effectiveness of design
- Information on the effectiveness of management
- Information on the effectiveness of implementation
- Information on the effectiveness of the evaluation itself
- Information on the achievement of objectives
- Information on the broader impacts of the instrument

Besides the production of information and the opportunity for learning, what other issues may be relevant for usefulness?

In some cases, it may be desirable to undertake an evaluation for a broader audience, such as the participants in a programme, or even potential participants, in order to generate support and a greater feeling of stakeholder value or to attract further participants. This broader dissemination may even include the general public, who as taxpayers may feel entitled to seek reassurance that their contributions are being appropriately used.

Similarly, to some extent, it may be necessary to conduct an evaluation in order to justify political goals. One notable example is in the case of Structural Funds where evaluation is a pre-condition of the award of such support. At the other end of the scale, political administrations may conduct evaluations in order to justify their political decision to support (or terminate) a specific policy instrument in the face of countervailing arguments from political oppositions.

However, in the following, we will limit much of our analyses and discussion to the issue of usefulness in connection with the process of learning how to improve the delivery of policy instruments. Here, we refer to delivery in its broadest sense, and include formulation, design, management and implementation.

3 How to increase the usefulness of an evaluation

There are a number of assumptions and guidelines that may be advanced on how the usefulness of an evaluation may be increased. These can provide the basis for a number of hypotheses which may be tested on our available datasets.

One assumption that may be made is that, the more extensive and rigorous an evaluation, the greater its usefulness. That is the, more questions posed, issues examined and methodologies applied, the greater will be the usefulness of an evaluation to those commissioning it. Of course, this will be susceptible to a law of diminishing returns: Miles and Cunningham (2006) stress that it is advisable to have “a clear hierarchy of information requirements in order that the collection of data is prioritised in strict accordance to its utility”.

A further means to increase the usefulness of an evaluation identified by Miles and Cunningham (2006) is by gaining the trust and compliance of all stakeholders. The implication is that stakeholders will then feel more a part of the process and will be inclined to impart more useful and meaningful information. Miles and Cunningham state that such compliance may be achieved and that implicit or explicit resistance to the evaluation process may be minimised through demonstrating the utility of the evaluation process to stakeholders. This argument exhibits a degree of tautology in that, for an evaluation to be more useful it must demonstrate its usefulness! Nevertheless, it is clear that a willing set of interviewees or survey respondents are far more likely to produce evidence that can serve the purposes of the evaluation better than similar sets that are apathetic or resistant.

At a finer level of detail, Miles and Cunningham also note that some degree of exposure of evaluation principles and demonstration of their utility is very important – especially in circumstances where influential policymakers are present. Thus, the utility of the evaluation (in terms of the acceptance of its findings) will be enhanced if the evaluation report provides a clear rationale for the selection of methodologies it uses.

In turn, however, this implies that, in order to benefit from an evaluation, programme managers and others who commission evaluations need to be informed clients. Again as noted by Miles and Cunningham, this means that they should themselves be fully conversant with the utility and demands of the evaluation process and, ideally, will have received some training in commissioning and even in undertaking evaluations. Thus they should be capable of specifying features of the evaluation process and as acting as informed clients who are able to assess the quality and significance of the evaluation report.

Miles and Cunningham note that “there is no ‘magic bullet’, i.e. a single methodology that is applicable to all programmes, and which yields information on programme performance without requiring (and hopefully contributing to) knowledge about the wider system in which the intervention operates. Such false expectations will limit the scope and utility of evaluation in innovation programmes…” One conclusion arising from this is that evaluations that employ a range of methodologies might be expected to provide a greater range of information on programme performance and, hence, will tend to be more useful. Of course, the choice of methodology(ies) is greatly dependent upon the purpose of the evaluation and its intended use and the extent to which further methodologies are applied must respond to the law of diminishing returns already noted above.

A further suggestion is that the setting of clear, verifiable and measurable objectives is a particularly useful exercise in the planning of an evaluation. Thus, it may be assumed that such a process would contribute to the utility of the evaluation by making a clear association between the programme objectives and the findings of the evaluation. As noted by Miles and Cunningham, “the setting of verifiable or measurable objectives is a useful task as it:

- Clarifies and makes explicit the link between the programme or activity level objectives and higher level policy aims and objectives.
- Creates a direct link between the problem to be addressed and the analysis of what needs to be done.
- Can help to provide a common understanding of what are the important aims and activities which can assist in the implementation of the programme.
- Provides a clear basis for the definition of indicators which may be used to measure progress and assess achievement.
- Lays the basis for the evaluation of the programme and assists the evaluators in determining the degree of success the programme has achieved.”

This last suggestion on the use of verifiable objectives implies that evaluation is a task that must be planned for within the overall policy cycle, rather than a ‘bolt-on’ exercise instigated once the programme is underway or even completed. In addition, planning of an evaluation must be distinguished from the simple routinisation of evaluation as just another task within the implementation of a programme, where it becomes reduced to the level of an audit rather than as an opportunity for policy learning.

4 Potential areas for analysis in the database

The next stage of this case study examined the database of template responses arising from the survey. Two main lines of analysis were followed:

- Evidence of utility provided by the responses
- Testing of hypothesised links between utility and other database variables

The analyses examined Dataset C (questionnaires which had been validated by policy makers/programme managers).

4.1 Proxies and other indicators of usefulness from the templates

The first issue was to look for database variables that might be used as some sort of proxy for the utility of the specific evaluation under review (the questionnaire did not actually directly ask if the appraisal had been useful, or to what degree: this would have required an input from the programme manager or other responsible official and would probably have generated a more nuanced answer rather than a simple ‘yes’ or ‘no’). In the absence of these proxies, further analysis would not be possible.

Clearly, the best indicator of usefulness captured by the template is provided in Questions E.2.a to E.2.h which sought information on whether the appraisal contained any recommendations and, if so, which areas they addressed and the extent to which the recommendations were useful. The specific areas/outcomes identified were:

1. Changes to the design of the programme/measure appraised (E.2.c)
2. Changes to the management and implementation of the programme/measure appraised (E.2.d)
3. Changes to the design, management and implementation of future programmes/measures (E.2.e)
4. Changes to the design, management and implementation of contemporaneous programmes/measures (E.2.f)
5. Changes to broader policy formulation and implementation (E.2.g), and
6. Other outcomes (E.2.h)

As can be seen below, these correlate closely to the issues highlighted in Section 2 as potential contributors to usefulness:

- Information on the effectiveness of design
- Information on the effectiveness of management
- Information on the effectiveness of implementation
- Information on the effectiveness of the evaluation itself
- Information on the achievement of objectives
- Information on the broader impacts of the instrument

For analytical purposes, it is possible to combine all these outcomes (1-6, above) into a broad 'usefulness' indicator, although it should be borne in mind that the question utilised a Likert scale for responses ranging from 'not at all' to 'extensive'. This suggests that responses scoring 3 or above only should be used as other responses could be interpreted as of no or little utility. However, as the analyses compared means or median values, this issue was not thought to be significant.

The analysis could be further elaborated by combining outcomes 1 and 2 into 'internal' (programme specific) outcomes and 3-5 as 'external' (broader policy) outcomes, with the same considerations made for the Likert categories noted above.

One major caveat associated with this proxy for usefulness is that the absence of change (i.e. no or marginal outcomes noted) does not imply that the appraisal was not useful. For example, the measure may have been positively evaluated and found to require no adjustment. Similarly, the evaluation may not have been intended to have outcomes on broader policy (external outcomes) although responses to these questions would have elicited a 'not applicable' response.

However, this aspect may be checked from the responses to Questions F.3.a to F.3.g which investigated the consequences of the appraisal. Possible outcomes listed are:

1. Termination of the measure (F.3.a)
2. Major re-design of the measure (F.3.b)
3. Minor re-design of the measure (F.3.c)
4. Expansion/Prolongation of the measure (F.3.d)
5. Re-design of another measure (F.3.e)
6. Merger of measures (F.3.f)

The template also asked "If there were no consequences, please say why not?"

Notification of any of these outcomes could thus be assumed to be an indication of the usefulness of the appraisal. However, it is important to recognise that the outcomes may have been predetermined prior to the evaluation and the usefulness of the appraisal may have been limited to a 'rubber-stamp' endorsement of this action.

A further indication of the usefulness of the appraisal is provided by the response to Question F.4.a: "What aspects/methods were particularly useful in this appraisal for both this measure and for general policy making?" While this is a rather specific question and does not investigate the overall utility of the evaluation, it does offer a further means of identifying whether or not the appraisal was useful and in what way.

Thus, in the following analyses we will use a composite proxy of usefulness compiled from the answers to Questions E.2.a to E.2.h. This composite indicator is further elaborated into 'internal' (programme specific) outcomes and 'external' (broader policy) outcomes.

The analysis on the perceived usefulness of recommendations has been done with data set C, which includes only templates that the relevant policy-makers have confirmed and corrected. As the usefulness questions (e2c to e2g) were formulated on a 5 point Likert scale, which ranges between "1- Not at all" to "5 - Extensive", medians rather than means for this ordinal data have been used. Similarly composite variables have been calculated from the joint medians of corresponding variables.

Furthermore, where appropriate, these medians have been averaged. Mann-Whitney U tests have been used to associate binary variables and usefulness, while for k-independent samples the equivalent Kruskal Wallis test has been employed. All tests have been conducted at the 90% level of confidence.

4.2 Testing of hypothesised links between utility and other database variables

Having arrived at a useful proxy of usefulness, we can now perform a number of analyses to investigate correlations between this characteristic and other variables from the templates. These are dealt with below.

4.2.1 Basic statistics from the analyses

Preliminary analysis of the data indicated that of 132 completed templates, 84.8% (112) noted that the appraisal had contained recommendations. As shown in Table 28 below, in 84 appraisals the recommendations related to the design of the programme/measure, in 86 appraisals they related to the management or implementation of the programme/measure, in 84 cases they related to the design, management and implementation of future programmes, in 62 cases the recommendations applied to the design, management and implementation of contemporaneous programmes and in 79 appraisals, the recommendation related to broader policy formulation and implementation.

Table 28: Appraisals containing recommendations and their policy target

	Design of the programme	Management and implementation of the programme	Design, management & implementation of future programmes	Design, management & implementation of contemporaneous programmes	Broader policy formulation and implementation
Reccs	84	86	84	62	79
Not valid	48	46	48	70	43
% of templates with recommendations	64%	65%	64%	47%	60%

An initial conclusion from this analysis is that recommendations relating to the design, management and implementation of future programmes were considered to be equally useful as those relating to both the design and the management and implementation of the measure under review. Thus, the appraisals in our sample clearly play a strong role in both the ‘internal’ policy cycle of the measure and in the wider ‘external’ policy cycle concerning future programmes.

When the results were treated to provide a composite indicator of usefulness (i.e. by combining all templates for which recommendations had been provided and were considered to be useful) the following results were obtained (Table 29). These have been further disaggregated into internal (pertaining to the appraised programme – i.e. first two policy targets) and external (pertaining to other programmes or general policy – i.e. third, fourth and fifth policy targets) usefulness indicators. Note that it is possible for the sample size of these composite indicators to exceed those of their components as it is not necessary for all of the policy targets to be addressed by the composite indicator).

Table 29 Usefulness composite indicators (median)

	N	Mean	Std. Deviation
Usefulness	110	3.30	1.038
Internal Usefulness	92	3.25	1.120
External Usefulness	94	3.11	1.218

These results seem to indicate that recommendations addressing internal usefulness (i.e. those which relate directly to the programme being appraised) are slightly more useful than those which address external usefulness (i.e. those which address other programmes (future or contemporaneous) or broader policy). All were significantly positively correlated.

4.2.2 Use of external evaluators

The assumption to be tested here is whether the use of external evaluators provides appraisals that are more useful than those that are conducted in-house. The analysis compared the medians of the composite indicators for usefulness between appraisals where external evaluators had been used and those where the appraisal had been conducted internally. The results are shown in Table 30.

Table 30: Usefulness and use of internal vs external evaluators

	Who conducted the appraisal					
	Internal		External		Mixed	
	Mean	Valid N	Mean	Valid N	Mean	Valid N
Usefulness	2.60	5	3.36	94	3.18	11
Internal Usefulness*	4.60	5	3.16	76	3.23	11
External Usefulness*	2.00	5	3.16	79	3.20	10

The results indicate that appraisals conducted by external or mixed evaluation teams had a higher level of external usefulness compared to those performed by internal teams, as might be expected. Conversely, internal usefulness was higher in appraisals conducted by internal teams than for appraisals conducted by external or mixed teams. Although the overall usefulness (external and internal) was also higher for appraisals conducted by external or mixed teams, this was not statistically significant.

4.2.3 Timing of evaluation

This analysis examined if any particular type of evaluation (*ex ante*, interim, accompanying or *ex post*) was more useful than other. The assumption is that *ex ante* evaluations might be more useful in an internal context (particularly regarding the design of the programme) whilst *ex post* evaluations could be expected to be of use in a broader sense with regard to other programmes and policy in general.

Table 31: Timing of appraisal and usefulness

	Timing of Appraisal									
	Ex-Ante		Interim		Accomp.		Ex-Post		Other	
	Mean	Valid N	Mean	Valid N	Mean	Valid N	Mean	Valid N	Mean	Valid N
Usefulness*	2.68	14	3.49	49	3.22	16	3.45	28	2.33	3
Internal Usefulness*	2.62	13	3.32	44	3.41	16	3.66	16	2.00	3
External Usefulness	2.25	8	3.30	44	2.93	15	3.25	24	2.33	3

The results (Table 31) indicate that interim and *ex post* evaluations have a slightly higher overall usefulness than other types of evaluations, while *ex post* evaluations have a higher internal usefulness than other types. No significant differences could be found for external usefulness. *Ex ante* evaluations tended to be significantly least useful in terms of internal and overall usefulness.

4.2.4 Purpose of evaluations: Summative versus formative

It might be hypothesised that formative evaluations are likely to be of greater utility than summative ones given the implications and definition of the term ‘formative’.

Table 32: Usefulness of summative and formative appraisals

	Purpose of Appraisal							
	Summative		Formative		Both		Other	
	Mean	Valid N	Mean	Valid N	Mean	Valid N	Mean	Valid N
Usefulness*	3.73	15	3.24	45	3.44	44	1.67	6
Internal Usefulness	3.50	9	3.23	42	3.29	35	2.83	6
External Usefulness*	3.70	15	3.09	38	3.11	35	1.67	6

Overall, the results (Table 32) show that summative evaluations have a higher overall usefulness and external usefulness compared to formative and combined approaches. No significant difference was detectable for internal usefulness, although the average median value was again higher for summative evaluations. This is an interesting outcome as one might anticipate formative evaluations to have a higher degree of utility. However, if the policy maker wished the evaluation to deliver evidence (on the programme performance) on which to base a formative judgement or to justify a decision, then a summative evaluation could provide this, i.e. it would be of use. There may be a definitional issue at stake here – how does one define a formative evaluation: is it one that is expected to deliver specific recommendations for change - is the formative nature of the evaluation contained within it or does it depend on the eventual consequences of the evaluation?

4.2.5 Evaluation of Structural Fund measures and portfolio evaluations

It is a condition of the Structural Funds that measures supported under them should be subject to an evaluation. Where such conditionality is introduced, it might be hypothesised that the appraisals undertaken would have less broad policy impact and that the results would apply more directly to the measures under consideration than to other future and contemporaneous measures (see also Section 4.2.6 below).

The opposite might be expected for portfolio evaluations, where a number of measures are appraised as a package under an umbrella evaluation and where broader policy lessons might be expected.

As can be seen from Table 33 below, the assumption regarding the usefulness of Structural Fund evaluations is fully supported: internal, external and overall usefulness are all significantly higher for non-Structural Fund appraisals than for those supported by the Structural Funds. The same trend is evident for portfolio evaluations, with these being considered significantly less useful than non-portfolio evaluations. The difference is statistically significant for all categories of usefulness.

Table 33: Usefulness of Structural Fund and portfolio evaluations

	Non-Structural Fund		**Structural Fund**		**Non-portfolio**		**Portfolio**	
	Mean	Valid N	Mean	Valid N	Mean	Valid N	Mean	Valid N
Usefulness	3.46	87	2.72	23	3.40	100	2.35	10
Internal Usefulness	3.40	70	2.77	22	3.34	82	2.55	10
External Usefulness	3.26	78	2.38	16	3.17	86	2.38	8

4.2.6 Conditional evaluations

This analysis is similar to the question posed in the preceding section on Structural Fund measures, but relates to a specific question in the template regarding whether the appraisal was subject to a conditional evaluation. Here the assumption is that conditional evaluations (i.e. are more likely to be reporting oriented) would tend to be more useful than those that are undertaken for the purpose of policy learning.

Table 34: Usefulness and conditionality

	Appraisal as a condition of external/international (co)sponsorship			
	No		Yes	
	Mean	Valid N	Mean	Valid N
Usefulness	3.40	68	2.65	30
Internal Usefulness	3.17	54	3.05	30
External Usefulness	3.26	59	2.13	23

In this case (Table 34) the overall and external usefulness were both significantly higher in cases where the appraisal was not conducted as a condition of the external or international funding. This supports the findings in 4.2.5 where the appraisals of measures supported by the Structural Funds were found to have a lower level of usefulness then measures that had been evaluated non-conditionally.

4.2.7 Planned evaluations

Again, this issue is somewhat related to that of conditional appraisals (4.2.6) and the existence of dedicated budgets for the evaluation (4.2.8). Here, it could be assumed that evaluations that are planned for in the design of the measure would be more useful than those that have been evaluated on a more *ad hoc* basis.

However, as can be seen from the results (Table 35), there was no significant difference between evaluations that had been planned for during the design stage and those that had not. This may not be as surprising as it initially seems since a planned evaluation may be one that is fully integrated within the policy cycle of programme design or it may simply be required as part of a routinised process of audit: in our analysis it was not possible to distinguish between these two extremes, which clearly have widely different consequences for the notion of utility.

Table 35: Planning and usefulness

	Appraisal foreseen and planned for			
	No/Don't Know		Yes	
	Mean	Valid N	Mean	Valid N
Usefulness	3.64	18	3.24	92
Internal Usefulness	3.50	12	3.21	80
External Usefulness	3.50	16	3.03	78

4.2.8 Dedicated budget for appraisals

Again related the question posed above, the existence of a dedicated budget for the appraisal might be an indication that it was planned in advance and had sufficient resources to meet its objectives.

Table 36: Dedicated budget and usefulness

	Dedicated Budget for Appraisal			
	No		Yes	
	Mean	Valid N	Mean	Valid N
Usefulness	3.22	41	3.33	47
Internal Usefulness	3.15	36	3.21	38
External Usefulness	3.06	35	3.05	42

However, as in the above Section relating to planning, there was no apparent relationship between the existence of a dedicated budget for the appraisal and its usefulness. Again, the same rationale as for planned evaluations may apply.

4.2.9 Topics and utility

This set of analyses was more investigative, with no prior assumptions and looked at whether any specific topics covered by the appraisal could be associated with its utility.

Table 37: Topics covered by the appraisal and usefulness (1)

	External Consistency				Internal Consistency				Coherence/Complementarity			
	Not Covered		Covered		Not Covered		Covered		Not Covered		Covered	
	Mean	Valid N	Mean	Valid N	Mean	Valid N	Mean	Valid N	Mean	Valid N	Mean	Valid N
Usefulness	3.40	20	3.28	89	3.36	18	3.29	91	3.20	32	3.25	73
Internal Usefulness	3.50	15	3.20	76	3.75	14	3.18	77	3.42	25	3.06	62
External Usefulness	3.11	19	3.11	74	3.06	16	3.12	78	2.86	28	3.12	62

In terms of external consistency (Table 37), there was no significant difference in usefulness between evaluations that covered this aspect and those that did not.

With regard to internal consistency, the internal usefulness of appraisals where this topic was NOT covered tended to be higher than those where it was covered (but not significantly); no real difference was observed for overall and external usefulness.

Evaluations which considered coherence/complementarity showed mixed results for the categories of usefulness, although none of these were significantly higher or lower. However, one of the components of external usefulness – ‘changes to the design, management and implementation of future programmes/measures’ was significantly higher in cases where the issue was covered.

From Table 38 it may be seen that overall and external usefulness were significantly higher in evaluations that had looked at goal attainment/effectiveness. While internal usefulness was lower in cases where this aspect was not covered, the difference was not significant.

Table 38 Topics covered by the appraisal and usefulness (2)

	Goal Attainment/Effectiveness				**Outputs, Outcomes and Impacts**			
	Not Covered		Covered		Not Covered		Covered	
	Mean	Valid N	Mean	Valid N	Mean	Valid N	Mean	Valid N
Usefulness	2.96	14	3.36	90	3.14	11	3.29	97
Internal Usefulness	3.63	12	3.18	74	3.50	10	3.18	80
External Usefulness	2.50	10	3.14	80	2.75	8	3.10	84

The coverage of outputs, outcomes and impacts made no difference to the overall usefulness of appraisals. However, external usefulness was higher where these were covered whilst the reverse was true for internal usefulness. None of the results were significant.

The usefulness of appraisals in which the quality of outputs was examined showed some variation but no significant differences in terms of usefulness (Table 39).

Table 39: Topics covered by the appraisal and usefulness (3)

	Quality of Outputs				**Value for Money/Return on Investment/Cost-Benefit Efficiency**			
	Not Covered		Covered		Not Covered		Covered	
	Mean	Valid N	Mean	Valid N	Mean	Valid N	Mean	Valid N
Usefulness	3.21	47	3.29	55	3.25	73	3.24	25
Internal Usefulness	3.32	41	3.07	43	3.24	60	3.11	23
External Usefulness	2.95	37	3.13	51	3.01	65	3.05	20

Appraisals that looked at Value for Money, Return on Investment and Cost-Benefit Efficiency displayed very little difference in all levels of usefulness. However, one of the external usefulness factors – ‘Changes to the design, management and implementation of contemporaneous programmes/measures’ showed a significantly higher result for evaluations where it was NOT covered.

Table 40: Topics covered by the appraisal and usefulness (4)

	Programme Implementation Efficiency				Project Implementation Efficiency			
	Not Covered		Covered		Not Covered		Covered	
	Mean	Valid N	Mean	Valid N	Mean	Valid N	Mean	Valid N
Usefulness	3.11	19	3.35	84	3.20	58	3.32	41
Internal Usefulness	3.03	15	3.29	72	3.19	50	3.28	34
External Usefulness	2.94	18	3.15	72	2.90	47	3.21	39

The results presented in Table 40 (above) indicate that there was little difference between the usefulness of appraisals that examined Programme Implementation Efficiency. However, as might be expected, 'Changes to the design, management and implementation of future programmes/measures' – one of the internal usefulness components, was significantly higher in evaluations that covered this aspect.

Examination of Project Implementation Efficiency appeared to have little effect on usefulness for all categories, although external usefulness was slightly (not significantly) higher for evaluations where this aspect was covered.

Appraisals which examined Input Additionality (Table 41 – below) showed the unusual result that where these topics were covered, they were considered less useful than where they were not in terms of internal and overall usefulness. Only the former result was significant, however.

No statistically different differences were found for evaluations which used Output Additionality and those which did not.

Table 41: Topics covered by the appraisal and usefulness (5)

	Input Additionality				Output Additionality			
	Not Covered		Covered		Not Covered		Covered	
	Mean	Valid N	Mean	Valid N	Mean	Valid N	Mean	Valid N
Usefulness	3.32	55	3.11	41	3.18	47	3.37	51
Internal Usefulness	3.40	51	2.92	33	3.39	41	3.16	45
External Usefulness	3.01	47	3.00	35	2.89	40	3.18	46

Mixed results were obtained for evaluations which examined Behavioural Additionality: those which covered the topic were more externally useful than those which did not, whilst the converse applied for internal usefulness (Table 42). All differences were not statistically significant.

Table 42: Topics covered by the appraisal and usefulness (6)

	Behavioural Additionality				Policy/Strategy Development			
	Not Covered		Covered		Not Covered		Covered	
	Mean	Valid N	Mean	Valid N	Mean	Valid N	Mean	Valid N
Usefulness	3.18	47	3.37	51	2.98	24	3.42	84
Internal Usefulness	3.39	41	3.16	45	3.40	21	3.24	70
External Usefulness	2.89	40	3.18	46	2.58	18	3.25	74

However, appraisals that examined Policy or Strategy Development were found to be statistically more useful (in terms of overall and external utility) than those which did not cover this aspect, as might be anticipated. The reverse was true for internal utility, although this was not a statistically significant result.

Table 43: Topics covered by the appraisal and usefulness (7)

	Gender issues				**Minority issues**			
	Not Covered		Covered		Not Covered		Covered	
	Mean	Valid N	Mean	Valid N	Mean	Valid N	Mean	Valid N
Usefulness	3.33	72	3.05	28	3.19	88	3.00	8
Internal Usefulness	3.30	62	3.10	26	3.18	76	2.94	8
External Usefulness	3.06	62	2.98	23	2.94	88	3.20	5

With regard to the final two topics, Gender Issues and Minority Issues (Table 43), the usefulness of appraisals that did not consider these issues was very slightly higher than those in which they were covered – with the exception of external usefulness in the case of minority issues - but none of these results were statistically significant).

In summary, the only logical (i.e. anticipated) relationship from the above set of analyses were the greater overall and external usefulness associated with evaluations that considered Goal Attainment/Effectiveness and the greater degree of overall and external usefulness of appraisals which examined Policy and Strategy Development. There does not appear to be a clear explanation for the general trend where usefulness was often higher in cases where specific topics were not covered.

4.2.10 Impacts examined by the appraisal and usefulness

This set of analyses was also investigative and sought to find any links or relationships between usefulness and the types of impact (Scientific, Technological, Economic, Social, Environmental, etc.) examined by the appraisal.

From Table 44, it appears that the external usefulness of appraisals which examined the scientific impact on the participants and beyond was higher than other cases. This result was statistically significant.

In terms of technological impact, the overall usefulness and external usefulness of appraisals which examined the technological impact on the participants and beyond was greater (significantly in the case of external usefulness). The reverse was the case for internal usefulness, but not significantly.

Table 44: Types of impact examined and usefulness (1)

	Scientific Impact						Technological Impact					
	No		Impact on participant		impact on participants and beyond		No		Impact on participant		impact on participants and beyond	
	Mean	Valid N	Mean	Valid N	Mean	Valid N	Mean	Valid N	Mean	Valid N	Mean	Valid N
Usefulness	3.30	59	2.90	15	3.54	28	3.27	41	3.08	24	3.51	36
Internal Usefulness	3.34	50	2.93	14	3.25	20	3.42	33	3.16	22	3.22	30
External Usefulness	2.98	49	2.93	14	3.56	25	2.93	37	2.66	19	3.67	29

Appraisals that looked at the economic impact on the participants tended to have a higher overall usefulness than those that looked at the impact on the participants or beyond, or which did not cover this impact (Table 45). External usefulness appeared to be higher in evaluations that looked at either the impact on the participants or on the participants and beyond. None of these relationships was significant, however.

From the same Table 45, it can be seen that the effects of looking at the Social Impact on various groups of participant was similar except that all categories of usefulness were higher for appraisals which examined impact on the participants only. Again, no relationships were statistically significant.

Table 45: Types of impact examined and usefulness (2)

	Economic Impact						Social Impact					
	No		Impact on participant		impact on participants and beyond		No		Impact on participant		Impact on participants and beyond	
	Mean	Valid N	Mean	Valid N	Mean	Valid N	Mean	Valid N	Mean	Valid N	Mean	Valid N
Usefulness	3.22	32	3.50	28	3.26	44	3.17	54	3.75	4	3.28	39
Internal Usefulness	3.45	30	3.40	24	3.07	34	3.34	48	3.50	3	2.92	30
External Usefulness	2.84	28	3.22	23	3.27	37	2.91	46	3.25	4	3.17	32

Finally, there was a general tendency for evaluations that did NOT examine the Environmental impact of measures (Table 46) to have a slightly higher level of overall, internal and external usefulness. Only the result for internal usefulness was statistically significant. This finding may be linked to the requirement of Structural Fund evaluations to include an examination of environmental impact – as noted above, such evaluations tended to have lower usefulness ratings.

Table 46: Types of impact examined and usefulness (3)

	Environmental Impact					
	No		Impact on participant		impact on participants and beyond	
	Mean	Valid N	Mean	Valid N	Mean	Valid N
Usefulness	3.37	64	3.25	4	2.70	20
Internal Usefulness	3.44	54	3.00	4	2.47	17
External Usefulness	3.13	56	3.00	3	2.59	17

4.2.11 Sponsor effects

It might be assumed that appraisals sponsored by the programme's management, as opposed to other government departments or public bodies could affect their utility. Thus, this question examined the affects of the sponsor on usefulness. The outcomes are shown in Table 47.

Table 47: Appraisal Sponsor and usefulness

	Programme Owner / Manager				Other Government Department				Other Public Bodies			
	No		Yes		No		Yes		No		Yes	
	Mean	Valid N	Mean	Valid N	Mean	Valid N	Mean	Valid N	Mean	Valid N	Mean	Valid N
Usefulness	2.95	11	3.36	98	3.01	66	3.52	21	3.04	77	3.86	7
Internal Usefulness	2.85	10	3.32	81	3.13	55	3.32	19	3.08	66	4.20	5
External Usefulness	2.29	7	3.19	86	2.72	53	3.21	19	2.73	64	4.20	5

Overall, it appears that appraisals sponsored by the programme owner or manager were more useful at all levels than those that were not sponsored by the programme owner or manager. In the case of external usefulness, this result was statistically significant. The same was true for appraisals that were sponsored by other Government departments (with the results for overall and external usefulness being statistically significant) and (with the same statistically significant results) for those sponsored by other public bodies. .

4.2.12 Availability and language of report

This analysis examined whether the level of availability of the appraisal report or the language in which it was written, had any influence on its utility. The results are presented below (Table 48 and Table 49).

Table 48: Level of availability of appraisal report and utility (1)

	Availability of the Appraisal Report							
	Obtainable on request		Internal use only		Published (Hard Copy)		Published (Web)	
	Mean	Valid N	Mean	Valid N	Mean	Valid N	Mean	Valid N
Usefulness	3.50	4	3.50	10	3.61	9	3.24	85
Internal Usefulness	3.50	3	3.25	8	3.44	8	3.22	71
External Usefulness	3.33	3	3.56	8	3.71	7	2.99	74

Interestingly, the results (Table 49) seem to indicate that appraisals that were published in hard copy scored higher on overall usefulness and external usefulness. However, it should be noted that these results were not statistically significant.

From Table 49 it may be deduced evaluation reports published in English (which may also have been the native language) had higher levels of usefulness (for all categories) than the other options. The result for external usefulness was statistically significant.

Table 49: Language of appraisal report and utility (2)

	Language of the Appraisal Report					
	Both		English		Native language	
	Mean	Valid N	Mean	Valid N	Mean	Valid N
Usefulness	3.06	8	3.57	21	3.26	81
Internal Usefulness	3.19	8	3.56	17	3.18	67
External Usefulness	2.33	6	3.53	17	3.11	94

4.2.13 Methodological effects on utility

The next set of analyses sought to detect any relationships between the usefulness of the appraisals and the data analysis and data collection methodologies used. The results are shown in Table 50-Table 55(for data analysis methods) and 29-33 (for data collection/source methodologies).

Table 50: Data analysis methodologies and usefulness (1)

	Case Study Analysis				Network Analysis			
	No		Yes		No		Yes	
	Mean	Valid N	Mean	Valid N	Mean	Valid N	Mean	Valid N
Usefulness	3.11	64	3.50	44	3.26	80	3.36	18
Internal Usefulness	2.99	55	3.64	35	3.24	70	3.09	11
External Usefulness	2.97	53	3.29	40	3.08	68	3.18	17

The results indicated that overall usefulness, internal usefulness and external usefulness were all higher, the former two significantly so, for evaluations that had used case study analysis (Table 50).

Appraisals that had used network analysis (of which there were relatively few), were found to have a slightly higher level of overall and external usefulness (Table 50). However, the trend was reversed for internal usefulness. In no instance was the relationship significant, however.

Table 51: Data analysis methodologies and usefulness (2)

	Econometric Analysis				Descriptive Statistics			
	No		Yes		No		Yes	
	Mean	Valid N	Mean	Valid N	Mean	Valid N	Mean	Valid N
Usefulness	3.26	83	3.03	16	3.32	25	3.22	81
Internal Usefulness	3.20	68	2.81	13	3.40	15	3.12	73
External Usefulness	3.04	71	3.07	15	3.28	20	2.97	70

The results for appraisals using Econometric Analysis (Table 51) were mixed with overall usefulness and internal usefulness being higher in cases where the methodology was not used, although neither result was statistically significant. External usefulness showed no difference.

Interestingly, there was a tendency for evaluations which did not use descriptive statistics (which is a very basic methodology, common to the majority of appraisals) to be rated of more use than those which did use descriptive statistics. None of these results were significant, however (Table 51).

There was a statistically significant difference between appraisals that used Input/Output Analysis and those which did not (Table 52), for all categories of usefulness. However, those appraisals which did NOT use this approach attracted a higher usefulness rating those which did.

Table 52: Data analysis methodologies and usefulness (3)

	Input/Output Analysis			
	No		Yes	
	Mean	Valid N	Mean	Valid N
Usefulness	3.34	80	2.80	22
Internal Usefulness	3.33	66	2.67	21
External Usefulness	3.14	69	2.47	17

Although the overall usefulness and external usefulness of evaluations that employed Document Analysis were higher than for those which did not (Table 53), the difference was not statistically significant. However, one of the components of external usefulness – changes to the design, management and implementation of contemporaneous programmes – did score significantly higher for evaluations where the methodology was used.

Table 53: Data analysis methodologies and usefulness (4)

	Document Analysis				**Context Analysis**			
	No		Yes		No		Yes	
	Mean	Valid N	Mean	Valid N	Mean	Valid N	Mean	Valid N
Usefulness	3.13	38	3.34	69	3.46	34	3.12	68
Internal Usefulness	3.23	33	3.19	56	3.43	29	3.02	55
External Usefulness	2.89	36	3.18	56	3.07	29	3.03	58

There was a significant negative link between the use of Context Analysis and internal usefulness (Table 53). A similar (but not significant) link was apparent for overall usefulness also.

There was a negative relationship between appraisals using Before/After Group Comparisons and all categories of usefulness (Table 54). This relationship was statistically significant for overall usefulness only.

Table 54: Data analysis methodologies and usefulness (5)

	Before/After Group Comparison Approach				Control Group Approach			
	No		Yes		No		Yes	
	Mean	Valid N	Mean	Valid N	Mean	Valid N	Mean	Valid N
Usefulness	3.31	89	2.64	11	3.24	84	3.33	15
Internal Usefulness	3.22	71	2.77	11	3.18	72	3.17	9
External Usefulness	3.12	77	2.60	10	3.02	73	3.38	13

No clear pattern emerged for Control Group Approaches and usefulness, although overall usefulness and external usefulness scores for appraisals using this approach were higher, but not significantly so (Table 54).

Despite a tendency for appraisals using Counter-Factual approaches to score higher on all categories of usefulness, none of the results were statistically significant (Table 55).

Table 55: Data analysis methodologies and usefulness (6)

	Counter-Factual Approach				Cost / Benefit Approach			
	No		Yes		No		Yes	
	Mean	Valid N	Mean	Valid N	Mean	Valid N	Mean	Valid N
Usefulness	3.24	77	3.48	20	3.40	81	2.58	19
Internal Usefulness	3.20	66	3.35	13	3.29	63	2.71	19
External Usefulness	3.08	67	3.18	17	3.22	72	2.33	15

In another interesting result, there was a strong statistically significant negative relationship between all categories of usefulness and the use of Cost/Benefit Approaches (Table 55).

In another set of rather interesting results, evaluations that did NOT use existing databases and surveys were rated (strongly) significantly more useful than those which did for all categories of usefulness (Table 56).

Table 56: Data collection methodologies/sources and usefulness (1)

	Existing Surveys / Databases				Participant Surveys			
	No		Yes		No		Yes	
	Mean	Valid N	Mean	Valid N	Mean	Valid N	Mean	Valid N
Usefulness	3.70	33	3.03	73	2.89	27	3.41	74
Internal Usefulness	3.79	24	2.94	64	3.08	25	3.25	58
External Usefulness	3.40	31	2.85	59	2.67	21	3.20	66

As can also be seen from Table 56, evaluations which used participant surveys were found to be rated more useful than those that did not, in all categories of usefulness. This result was significant for both overall usefulness and external usefulness.

No clear overall relationship could be detected for non-participant surveys, with overall usefulness and external usefulness scoring higher in evaluations which used these approaches, and internal usefulness scoring lower (Table 57). No result was statistically significant.

Table 57: Data collection methodologies/sources and usefulness (2)

	Non-Participant Surveys				Interviews			
	No		Yes		No		Yes	
	Mean	Valid N	Mean	Valid N	Mean	Valid N	Mean	Valid N
Usefulness	3.22	74	3.33	23	2.41	16	3.49	90
Internal Usefulness	3.21	64	3.03	15	2.86	14	3.36	74
External Usefulness	3.04	65	3.21	19	2.09	11	3.28	81

Evaluations using interviews (a commonly applied methodology) had a significantly higher usefulness rating for all categories of usefulness, with overall and external usefulness being strongly statistically significant (Table 57).

There was also an indication that appraisals using focus groups, workshops and similar approaches had a higher usefulness rating (for all categories) than those which did not employ these methodologies. This difference was significant for overall and external usefulness (Table 58).

Table 58: Data collection methodologies/sources and usefulness (3)

	Focus Groups / Workshops / Meetings				Peer Review			
	No		Yes		No		Yes	
	Mean	Valid N	Mean	Valid N	Mean	Valid N	Mean	Valid N
Usefulness	3.04	45	3.50	60	3.10	79	4.00	21
Internal Usefulness	3.11	38	3.36	49	3.07	68	4.00	14
External Usefulness	2.86	38	3.31	54	2.88	66	3.86	21

The anticipated utility of Peer Review was strongly supported by the results (Table 59) where there was a strong statistically significant relationship between all categories of usefulness and appraisals which employed this approach.

Although there was an apparently strong relationship between all categories of usefulness and the use of technometrics and/or bibliometrics in appraisals, this was not statistically significant – and was probably an artefact of the very low number of cases (Table 59).

Table 59: Data collection methodologies/sources and usefulness (4)

	Technometrics / Bibliometrics Search				Document Search			
	No		Yes		No		Yes	
	Mean	Valid N	Mean	Valid N	Mean	Valid N	Mean	Valid N
Usefulness	3.23	97	4.00	1	3.29	33	3.26	74
Internal Usefulness	3.15	79	3.50	1	3.48	26	3.09	63
External Usefulness	3.05	84	4.00	1	3.08	32	3.08	60

Also from Table 59, it can be seen that there was no clear relationship between usefulness and the use of document searches and usefulness.

Lastly in this set of analyses, while the use of monitoring data seemed to have no effect on overall and internal usefulness (Table 60), it did have a non-significant positive relationship with external usefulness. This latter was explained by a strong statistically significant relationship between the use of such data and the component – ‘Changes to broader policy formulation and implementation’.

Table 60: Data collection methodologies/sources and usefulness (5)

	Monitoring Data			
	No		Yes	
	Mean	Valid N	Mean	Valid N
Usefulness	3.22	18	3.23	83
Internal Usefulness	3.14	14	3.20	73
External Usefulness	2.73	15	3.11	70

4.2.14 Audience effects

The following set of analyses looks at the effects of the main intended audience on usefulness (see Table 61-Table 65).

Table 61: Main intended audience for the appraisal and usefulness (1)

	Policy Maker (Politicians)				Policy Maker (Government Officials)			
	No		Yes		No		Yes	
	Mean	Valid N	Mean	Valid N	Mean	Valid N	Mean	Valid N
Usefulness	3.29	42	3.33	59	2.83	3	3.26	101
Internal Usefulness	3.24	36	3.28	48	2.67	3	3.20	101
External Usefulness	3.09	38	3.10	49	3.00	3	3.05	86

From Table 61 it can be seen that no relationship with any category of utility was discernible where the main intended audience was policy makers (politicians). Likewise, there was no significant relationship between utility and appraisals intended for Government Officials, although there was a slight positive relationship for all categories of usefulness. It should be noted that the number of cases listed under the ‘No’ category was very low: most evaluations would, of course, be expected to be aimed at policy makers.

Table 62: Main intended audience for the appraisal and usefulness (2)

	Programme Management				Auditors / Financial Authorities			
	No		Yes		No		Yes	
	Mean	Valid N	Mean	Valid N	Mean	Valid N	Mean	Valid N
Usefulness	4.00	2	3.29	106	3.35	47	3.19	57
Internal Usefulness	3.50	1	3.24	89	3.54	39	2.88	47
External Usefulness	4.00	2	3.09	91	3.07	41	3.09	49

Rather interestingly, evaluations aimed primarily at the programme's management (Table 62) were apparently less useful than those not aimed at this audience (although not significantly). Again, however, the number of 'No' cases was very low and this result is again probably an artefact.

Also from Table 62, the internal usefulness of appraisals intended for auditors and financial authorities was significantly lower than for other audiences, whilst overall usefulness was lower (but not significantly). There was no strong relationship between this factor and external usefulness.

No clear relationship was evident between overall usefulness and appraisals aimed at those directly supported by the measure, although there was a (non-significant) negative relationship between this factor and internal usefulness and a (non-significant) positive relationship between it and external usefulness (Table 63).

Table 63: Main intended audience for the appraisal and usefulness (3)

	Those directly supported by the measure				External / International (co)-sponsors			
	No		Yes		No		Yes	
	Mean	Valid N	Mean	Valid N	Mean	Valid N	Mean	Valid N
Usefulness	3.28	52	3.34	52	3.38	59	3.18	45
Internal Usefulness	3.35	43	3.14	43	3.38	47	3.06	39
External Usefulness	2.98	41	3.26	50	3.12	54	3.06	36

Where the main audience was external or international (co)-sponsors, all forms of usefulness were lower, but with no significant results (Table 63).

Table 64: Main intended audience for the appraisal and usefulness (4)

	Potential users of the measure				Policy analysts			
	No		Yes		No		Yes	
	Mean	Valid N	Mean	Valid N	Mean	Valid N	Mean	Valid N
Usefulness	3.36	59	3.28	41	3.24	54	3.30	49
Internal Usefulness	3.37	47	3.11	35	3.31	45	3.06	40
External Usefulness	3.17	49	3.14	39	2.93	47	3.24	42

There was also a negative relationship between usefulness and cases where the main intended audience consisted of potential users of the measure, but again with no statistically significant results (Table 64). No clear overall relationship with usefulness could be detected in the case where the main intended audience was policy analysts.

Lastly (Table 65 below), there was a slightly higher level of overall and external usefulness associated with appraisals intended for the general public as the main audience, although this was not significant in either case.

Table 65: Main intended audience for the appraisal and usefulness (5)

	General Public			
	No		Yes	
	Mean	Valid N	Mean	Valid N
Usefulness	3.25	65	3.42	30
Internal Usefulness	3.20	53	3.23	24
External Usefulness	3.04	56	3.35	26

4.2.15 Breadth of discussion and usefulness

It was hypothesised that there may be a relationship between usefulness and the extent to which the appraisal report is discussed (outside the immediate target audience, which is generally programme managers).

Table 66: Breadth of discussion and usefulness

	Discussion			
	Within Government circles		With wider participants/stakeholders	
	Corr. coeff	Valid N	Corr. coeff.	Valid N
Usefulness	.183	84	.218	89
Internal Usefulness	.244	68	.313	73
External Usefulness	.137	74	.171	79

The analysis (see Table 66) indicated that there was a statistically significant relationship between overall and internal usefulness where appraisals had been discussed within government circles.

Similarly, the analysis also indicated that appraisals that had been discussed with wider participants and stakeholders had higher levels of overall and internal usefulness than those which had not.

4.2.16 Tender procedures and usefulness

The degree of openness in the tender procedure might be expected to have some relationship to the use of internal or external evaluators (see Section 4.2.2).

As can be seen from Table 67, appraisals which had resulted from internal or closed tender procedures were associated with in significantly lower overall and external usefulness ratings. Open tenders resulting in generally the most useful appraisals (the number of cases of ‘other’ tender approaches was too few to be statistically valid).

Table 67: Tender procedures used and usefulness

	Tender procedure used									
	Internal		No tender		Closed		Open		Other	
	Mean	Valid N	Mean	Valid N	Mean	Valid N	Mean	Valid N	Mean	Valid N
Usefulness	2.88	8	3.14	14	2.96	23	3.50	48	4.00	3
Internal Usefulness	2.71	7	3.54	13	3.05	21	3.29	39	3.50	1
External Usefulness	3.00	8	2.65	13	2.67	18	3.34	41	4.00	3

4.2.17 Consequences of the appraisal and usefulness

Clearly, the extent to which the appraisal had any consequences on the measure under review or on other measures would be expected to have an impact on usefulness. The analyses are presented in Table 68-Table 70 below.

In the few cases that the appraisal had resulted in the termination of the measure or programme (Table 68), there was a slight tendency for the internal usefulness to be higher (as would be expected), although the low number of cases led to the result not being significant.

Likewise, although once again the number of cases where the appraisal had resulted in a major redesign of the measure was low, all categories of usefulness were higher but none of these results were statistically significant (Table 68).

Table 68: Consequences of appraisal and usefulness (1)

	Termination of measure				Major re-design of measure			
	No/Don’t know		Yes		No/Don’t know		Yes	
	Mean	Valid N	Mean	Valid N	Mean	Valid N	Mean	Valid N
Usefulness	3.32	106	3.00	4	3.26	99	3.68	11
Internal Usefulness	3.24	88	3.50	4	3.24	82	3.30	10
External Usefulness	3.14	90	2.25	4	3.07	83	3.36	11

Although many more appraisals led to a minor re-design of the measure, the results (Table 69) did not match those reported above for major re-design: internal usefulness was significantly higher while external usefulness was actually statistically significantly lower.

Table 69: Consequences of appraisal and usefulness (2)

	Minor re-design of measure				Expansion/prolongation of measure			
	No/don’t know		Yes		No/Don’t know		Yes	
	Mean	Valid N	Mean	Valid N	Mean	Valid N	Mean	Valid N
Usefulness	3.36	54	3.25	56	3.08	70	3.70	40
Internal Usefulness	2.95	37	3.45	55	3.02	54	3.58	38
External Usefulness	3.37	49	2.82	45	2.91	58	3.42	36

As might be anticipated, the analyses for appraisals that had resulted in the expansion or prolongation of the measure (Table 69) were more consistent, with all categories of usefulness being statistically significantly higher in all such cases.

As might be expected, evaluations that had led to the re-design of other measures (Table 70) had a significantly lower level of internal usefulness than other evaluations. Although the logical corollary obtained and external usefulness was higher (together with overall usefulness), neither of these positive relationships were statistically significant.

Table 70: Consequences of appraisal and usefulness (3)

	Re-design of other measures				Merger of measures			
	No/Don’t know		Yes		No/Don’t know		Yes	
	Mean	Valid N	Mean	Valid N	Mean	Valid N	Mean	Valid N
Usefulness	3.26	88	3.48	22	3.28	96	3.46	14
Internal Usefulness	3.33	82	2.60	10	3.29	80	2.96	12
External Usefulness	3.01	72	3.43	22	3.06	81	3.42	13

Finally, as again might be expected, evaluations which led to the merger of measures (Table 70) had a higher level of external usefulness with the reverse being the case for internal usefulness (although neither was to a statistically significant level). Overall usefulness was slightly higher for this outcome.

4.2.18 Country-specific analyses

An analysis was performed on the level of usefulness attributed to evaluations according to the country in which the innovation support measure was implemented. The results are shown below (Table 71) for countries where the number of cases (i.e. relevant evaluations) was five or over.

Table 71: Country specific analysis

Country	Internal usefulness		External usefulness		Usefulness	
	Mean	Valid N	Mean	Valid N	Mean	Valid N
Austria	2.98	25	2.85	27	3.08	30
Czech Republic	1.75	6	1.67	6	1.67	6
Germany	3.36	11	4.04	13	3.92	13
Finland	-	-	4.00	7	4.00	7
Greece	3.08	12	2.43	7	3.00	12
Netherlands	3.21	7	3.60	5	3.56	9
Sweden	4.67	6	4.25	6	4.67	6
United Kingdom	3.40	5	2.67	3	3.20	5

Owing to the methodology used, where evaluations from different countries were assessed by separate policy analysts, it is not really valid to make inter-country comparisons. However, examination of the table provides some interesting observations on the relative intra-country rating of internal and external usefulness. Thus, external usefulness was more highly rated in Germany and the Netherlands, whilst internal usefulness was more highly rated in Austria (but marginally), the Czech Republic (again marginally), Greece, Sweden and the UK.

4.2.19 Usefulness and type of innovation support intervention

The final analysis examined whether the usefulness of the evaluation might be associated with any particular type or mode of innovation support measure. The categorisation of measures used was:

- Indirect measures (tax, etc.)
- Direct financial support for innovation activities
- Innovation management support and dissemination, innovation culture
- Development and creation of intermediary bodies, agencies etc.
- Mobility of personnel
- Creation of start-ups and Spin-Offs
- Networks & Clusters, collaboration and Technology/Knowledge Transfer
- Science-Industry cooperation
- Support for the uptake and diffusion of innovation

The results (Table 72) indicate that evaluations of measures supporting and promoting science-industry cooperation were significantly more useful across all categories of usefulness. Evaluations of measures aimed at the creation of start-ups and spin-offs were also significantly useful, at the external and overall level of usefulness.

Table 72: Modality of policy measure and usefulness

Measure type	Internal usefulness		External usefulness		Usefulness	
	Mean	Valid N	Mean	Valid N	Mean	Valid N
Indirect measures	2.88	4	2.00	1	2.88	4
Direct financial support	3.56	39	3.23	43	3.50	50
Innov. Management support, etc.	3.09	23	3.06	25	3.20	27
Intermediary bodies	3.19	8	3.50	8	3.50	8
Mobility	2.94	8	2.67	6	2.94	8
Start-ups and spin-offs	3.58	6	4.00	5	3.75	8
Networks, clusters, T/KT	3.55	19	3.56	24	3.82	25
Science-industry cooperation	3.69	21	2.98	24	3.52	26
Support for uptake and diffusion	2.89	19	2.69	16	2.87	19

4.2.20 Usefulness and quality

In the same way that usefulness is a subjective property of evaluations, so too is the notion of quality. The survey questionnaire included a number of questions that could be used to derive a proxy indicator for quality.

These questions related to:

- If the evaluation addressed the Terms of Reference
- If the methods chosen satisfied the Terms of Reference/purpose of the appraisal
- Whether the analysis was clearly based on the data given
- If the conclusions were based on the analysis
- Whether the design of the evaluation appropriate given the objectives of the evaluation and the nature of the policy measure
- If the information sources used in the report were well documented and referenced
- Was the application of the qualitative methods satisfactory?
- Was the application of the quantitative methods satisfactory?
- If relevant, were the societal, institutional, policy and economic contexts of the measure examined and analysed in sufficient detail?

Again, responses were arranged according to a five-point Likert scale according to the perceived level of agreement with the statement.

Most of the quality aspects showed a significant correlation with each other, with the highest correlations observed between: analysis based on given data and conclusions based on analysis; appropriate design and conclusions based on analysis; application of quantitative methods and analysis based on given data; and appropriate design and analysis based on given data. Further analyses and cross-correlations of quality with other evaluation variables are presented in Chapter 3.

Error! Reference source not found. displays the correlation between the individual quality ndicator(s) and the aggregated usefulness indicators.

Overall, it appears that internal usefulness is little affected by the quality of the evaluation (only in the case of the design meeting the objectives of the evaluation). External usefulness is more linked with quality aspects (in five out of nine dimensions). This may be due to the need for evaluations that are expected to have broader policy implications to satisfy a higher level of quality criteria – i.e. evaluations whose impact is restricted to the target programme only, might have to satisfy less stringent quality criteria. Another, linked, explanation is that evaluation has both process and product benefits: the quality of the process (which is not really measured by our questions) will be more closely associated with internal impacts on the target programme whilst the quality of the product (i.e. the evaluation report itself, which are addressed by the dimensions in the template) will have implications for subsequent and other programmes. In other words, our quality dimensions mainly measure aspects relevant to external usefulness. Most of the total usefulness correlations may be explained by their dependence on the external usefulness results, but for the correlations with 'application of quantitative methods' and 'conclusions based on analysis', the result is probably explained as an artefact of the pairwise statistical methodology.

Table 73: Correlation coefficients between quality and usefulness indicators (Spearman (r), pairwise)

		Internal Usefulness	External Usefulness	Total Usefulness
Evaluation addresses Terms of Reference	r	.167	.300*	.238*
	N	62	64	76
Design of the evaluation appropriate	r	.281**	.433**	.454**
	N	86	86	102
Methods chosen satisfy Terms of Reference/purpose of the appraisal	r	.144	.107	.186
	N	71	74	86
Application of the qualitative methods satisfactory	r	.185	.304**	.306**
	N	81	81	97
Application of the quantitative methods satisfactory	r	.097	.199	.211*
	N	76	75	89
Information sources well documented and referenced	r	.074	.123	.154
	N	85	87	103
Analysis clearly based on the data given	r	.141	.313**	.332**
	N	86	88	104
Analysis covers the broader context sufficiently	r	.137	.240*	.212*
	N	88	86	102
Conclusions based on the analysis	r	.116	.165	.201*
	N	87	89	105

Thus, it appears that, to some extent there is a link between the usefulness and the quality of an evaluation. However, this is not an intuitive connection since an evaluation only needs to meet the minimum standards (in terms of quality criteria) to be acceptable to the programme manager's/policy maker's needs – it might be argued that, above this quality threshold, the returns on investment (in terms of resources, time, sophistication of approaches) begin to diminish.

5 Conclusion

At the outset, it should be noted that this study into the relationship between an evaluation's usefulness and the various aspects of the evaluation process, the evaluation outcomes and other characteristics, such as the type of measure under appraisal, was based on a number of broad preconceptions of how utility might be perceived and what might be the major determining factors and these have been investigated using the survey results. Nevertheless, the study was also intended to be, to a large extent, exploratory and to throw up as many questions as possible in order to explore the ways in which evaluation is of use to policy makers as a learning tool.

As a consequence, the results of the various analyses carried out present a very mixed picture. In summary, the major points are as shown below:

- 84% of evaluations examined had contained recommendations, with an almost equal balance between internal recommendations (relevant to the subject programme) and external recommendations (relevant to future programmes or to broader policy formulation).
- Evaluations addressing internal aspects of the programme had a slightly higher usefulness than those addressing external aspects.
- Significant positive correlations with at least one aspect of usefulness were identified for:
    - The use of an open tendering process when commissioning and evaluation
    - The use of external evaluators
    - The timing of the evaluation (*ex ante*, interim, *ex post*, etc.)
    - Summative over formative evaluations
    - Non-Structural Fund evaluations (i.e. a negative correlation between Structural Fund evaluations and utility)
    - Non-portfolio type evaluations (i.e. a negative correlation between portfolio type evaluations and utility)
    - Non-conditional evaluations (i.e. a negative correlation between conditional evaluations and utility)
    - Evaluations that examined the topics of goal attainment and effectiveness and policy/ strategy development
    - Evaluations that examined scientific impact and technological impact on the participants and beyond
    - Evaluations that employed case study analysis; participant surveys; interviews; focus groups/workshops and meetings; peer review
    - Evaluations that resulted in a minor redesign or expansion/prolongation of the measure
    - Evaluations sponsored by programme managers, other government departments or other public bodies
    - Evaluations not conducted primarily for auditors/financial authorities
    - Evaluations whose reports were published in English
    - Certain dimensions relating to the quality of the evaluation
- Negative correlations with at least one aspect of usefulness were observed for:
    - Evaluations that examined input additionality and environmental impacts
    - Evaluations that employed input/output analyses; context analysis; group comparison approaches; cost/benefit approaches; existing surveys and databases
- No significant correlations with any aspect of usefulness were detected for:
    - Evaluations planned during the design of the measure
    - Presence of a dedicated budget for the evaluation
    - Evaluations conducted primarily for policymakers (government officials) and programme management

    - Evaluations that examined outputs, outcomes and impacts; quality of outputs; value for money; programme/project implementation efficiency
    - Evaluations that employed monitoring data
    - Evaluations that had wider levels of availability
    - Evaluations where a major redesign of the measure resulted
- External usefulness was more highly rated in Germany and the Netherlands, whilst internal usefulness was more highly rated in Greece, Sweden and the UK
- The evaluations of measures for science-industry cooperation were significantly more useful across all categories of usefulness. Evaluations of measures aimed at the creation of start-ups and spin-offs were also significantly useful (external and overall).

Whilst a number of the statistically significant associations between usefulness and the survey variables were anticipated (based on our initial preconceptions), it is harder to explain some of the instances where negative correlations or where no correlations were detected. Several of the latter might be explained by the relatively low number of cases available within the analysis, whilst the prevalence of Structural Fund evaluations within the sample could also provide an explanation for the lack of, or negative correlations with usefulness, such as in relation to the examination of environmental impacts or where evaluations were planned for (both requirements of Structural Fund evaluations). Furthermore, nuances in the definition and interpretation of terms such as summative and formative, and ‘planned’ may also have played a role in negating some of our expected outcomes. All of these form important lessons for further studies and analysis of the issues under consideration.

In conclusion, the results of the analyses present a mixed picture, confirming some expectations yet failing to confirm or even refuting other expectations. As with most research endeavours, it is clear that further investigations are required into the aspect of usefulness and it is hoped that this study offers a valuable starting point – indeed, further analyses of the results obtained are possible and anticipated.

It must also be borne in mind that our analyses are based on proxy indicators of utility and are not based on the direct opinions of the programme managers concerned with each of the individual evaluation reports.

From a series of interviews with policymakers in the UK, it was noted that utility can be highly subjective (as was anticipated) and is primarily dependent on the overall purpose for which the evaluation was commissioned. Thus, evaluations may fulfil policy learning objectives, satisfy audit requirements, provide justification for a particular policy intervention, address specific sponsor needs, and so on. The primary criteria for usefulness, at least according to UK policymakers, were that an evaluation is deemed to be useful if the evaluation delivers the Terms of Reference in a consistent manner and if it provides actionable recommendations and delivers value for money. Usefulness can be defined as the degree to which there is feedback on policy and if something was learned from the process of the evaluation. However, the timing of an evaluation could have an impact on its usefulness – too early in the programme life cycle and there would be little to be learned, too late and it would not be possible to put the policy lessons into effect.

As a final point, this case study did not seek to look for extensive correlations between usefulness and quality (although this issue is examined briefly here and elsewhere in the report), largely since both were defined by proxy indicators in our analyses and it was felt that this would introduce too much variance to allow a clear interpretation of the results. As noted in the UK case study, the quality of an evaluation was defined by BIS policymakers as being fit for purpose, i.e. the evaluation meets the Terms of Reference within a reasonable budget and also delivers recommendations that are feasible and realistic. Thus, quality is to some extent very closely related to usefulness. Policymakers agreed that quality is an asymptotic function: that is there is a minimum level of quality that must be achieved for the delivery of the evaluation’s objectives – i.e. the programme manager who commissioned the evaluation must have confidence in the validity of the results and the recommendations. Any increase in the level of quality (i.e. through more complex data collection techniques, elaborate forms of analysis, etc,) incurs a law of diminishing returns (in terms of the usefulness of the evaluation). Thus, high quality evaluations would not necessarily imply a higher degree of usefulness.

Part II Chapter 5

The Role of Impact Assessment in Evaluation

This chapter looks at the role of impact assessments in evaluation. The study is based on a quantitative analysis of the INNO-Appraisal dataset and qualitative interviews with selected experts – policy makers, project officers and evaluators. The most important results are that impacts assessment is a typical element of nearly all evaluation studies considered, particular economic impacts and that there is a rather pragmatic use of impact assessments in terms of methods used (participant surveys and interviews). Nevertheless, we find that also elaborate methods which produce quantifiable results are used (econometric modelling, before-after comparison, control group approaches). Overall, impact assessments seem to try to respond to the demand for quantifiable results, despite methodological limitations or data availability problems. Policy makers are quite satisfied with impact assessments. Evaluations, which include impact analysis receive overall higher quality scores. For the future, we see a certain need to take into account more non-economic impact dimensions like gender issues, system effect and behavioural change.

Stephanie Daimer and Susanne Bührer, Fraunhofer ISI[68]

---

[68] The lead authors for this section of the report are as given above, with contributions from all other team members at different parts of its production, conclusions and interpretations have been approved and are shared by the group.

Table of Contents

Table of Contents 126

List of Tables 127

Table of Exhibits 127

Executive Summary 128

1 Introduction 131

  1.1 Data Sources used 132

    1.1.1 Descriptive Statistics 132

    1.1.2 Expert interviews 132

2 Relationship of Evaluation and Impact Assessment 133

  2.1 (Operational) Definitions 133

  2.2 Importance of Impact Assessments for Evaluation 133

  2.3 Characteristics of Evaluations with Impact Assessments 135

  2.4 Audiences of Evaluations with Impact Assessments 137

  2.5 The Use of Impact Assessments for certain types of policy measures 139

  2.6 Results from expert interviews: Main aim of evaluation studies compared to main purpose of impact studies 139

3 Performance of Impact Assessments 140

  3.1 Topics covered in Evaluations with Impact Assessments 140

  3.2 Methods used in Impact Assessments 141

  3.3 Results from expert interviews: Conceptual and methodological requirements for economic impact studies 143

4 The impact and usefulness of impact assessments 143

  4.1 Evidence from the evaluation database 143

  4.2 Results from expert interviews 146

    4.2.1 What kind of impact study is the most useful one for policy makers? 146

    4.2.2 Influence of impact studies in decision-making processes 146

5 Conclusions 147

References 150

List of Tables

Table 74: Economic impact coverage in relation to the number of impacts covered........................135
Table 75: Characteristics of Evaluations and the coverage of impact 136
Table 76: Topics covered in Evaluations compared to the coverage of impact 141
Table 77: Methods covered in Evaluations compared to the coverage of impact 142
Table 78: Quality of Evaluations with Impact Assessment 144

Table of Exhibits

Exhibit 40: Impact Dimensions covered by the evaluations 134
Exhibit 41: Addressed Audiences of Evaluations with Impact Assessments 138

Executive Summary

While how-to-do-guides set out frameworks from an (applied) scientific viewpoint, one might ask how impact assessments are being performed in reality – the reality of service contracts for programme owners, most probably including budget restrictions, customer needs, and tough schedules. One might ask whether there is some systematic use of methods, i.e. whether there are certain sets of methods which are employed for specific policy measures and in specific contexts. One might also ask whether evaluation studies of policy programmes have an impact on future innovation policy.

The quantitative analysis of the database shows a number of interesting results:

General results

- Impact assessment is a central function of evaluation studies: A large number of studies across Europe claim to do impact assessment, currently most important are economic impacts.
- Impact assessments appear to be **central and wide-spread** across Europe
- Impact studies of structural fund evaluations differ significantly from impact studies of national innovation programmes.

Impact types

- Typically we find the use of a very **broad definition** of impact assessment, including all types of effects
- Assessment of economic impact is most dominant, other impact types of importance are technological and societal impacts (not: scientific and environmental impacts)
- The assessment of **new impact types** (apart from economic or technological) is still rather seldom. Societal impacts are often covered with an estimation of new jobs having been created, but other topics, such as gender impacts are quite rare.
- A high number of evaluations claims to assess **indirect impacts**, i.e. spill-over effects beyond the participants of a programme. This is given the methodological difficulties for assessing economic or societal impacts a surprising result. This result seems to reflect the demand for results on these spill-over effects.
- **Additionality** concepts are well established beyond the UK. They are employed by half of the evaluations in the sample. This is also true for behavioral additionality which has obviously become an integral part of the idea of additionality.
- Structural fund evaluations more often cover social and environmental impacts.

Methods used

- Almost the whole toolbox of possible **methods** is employed for impact assessment, including elaborate methods such as a control group approach.
- Most of the impact assessments are qualitative and part of broader evaluation studies.
- There are only few quantitative impact assessments using elaborated quasi-experimental designs like control-group approaches.

- Impact assessment is typically not a mere econometric exercise, but **often used in a** contextually **sensitive way.**

Policy Cycle

- Impact assessment is not a clearly retrospective element of evaluation. Often, it is also used in the form of ex-ante impact assessment and in accompanying evaluations.
- Evaluations which include impact assessments, in particular the assessment of societal impacts, are more often used for external communication. Experts confirm that impact assessment is in particular important for legitimizing the political interventions.
- If impact assessments are included into evaluations this leads to higher quality scores.
- With respect to usefulness, evaluations of (single) national programmes seem to be more useful for policy makers than structural fund evaluations.

The analysis and the interviews indicate a set of clear recommendations. Most important issues from policy maker perspective are:

5) Impact assessments are an important part of evaluations, but often not the only one. It seems that impact assessments seem to make most sense within the framework of a "normal" evaluation study which for example covers context analysis extensively.
6) Evaluators have responded to the demand for quantitative results and employ a variety of (elaborate) methods to achieve them. However, in most cases it seems that the combination of qualitative and quantitative analysis can cope more adequately with impact assessments, as many impacts are not quantifiable at all.
7) Many pitfalls of impact studies can be avoided by a constant communication between policy makers and evaluators during the process of evaluation. This leads to transparency for the whole evaluation process in order to realize learning and to cope with methodological challenges.
8) As impact assessments clearly pursue the two purposes of learning and legitimation, two types of recommendations might be considered: Those designed for policy improvement implemented by the programme owners / managers and those directed to higher levels, which serve the legitimation aspect.

For the future, we think that it is useful to consider further impact dimensions to a greater extent than in the past. For example, economic Impact assessment is often intended, but not possible due to long-term effects and complex environments. Additionally we expect more mission oriented policy programmes where other topics like sustainability, customer needs and the structural / regional development might become more important. Looking at the demographic challenges in most European countries, the issue to integrate larger parts of society to the research sector will become even more relevant than in the past and therefore impact assessments should address gender and minority issues as well in more detail. Finally, the still prominent aspect of Behavioural Additionality in Innovation Programmes (e.g. innovation management, risk aversion) will remain important.

For impact assessment, this all means that it will become even more demanding to measure the intended effects – at least quantitatively. Given that non-economic impacts will gain more and more in importance this would mean that new sets of criteria and indicators will have to be defined, and most likely many of these indicators will be of a qualitative nature. So, quantitative method development is – although academically interesting – not the only promising way to go but again a method-mix between quantitative and qualitative approaches. More public support for experimental evaluation designs (including meta-evaluations at national as well as European level) could help to identify the most promising ways to identify new impact types.

However, impact assessment will even more than today require that from the complex set of programme goals one has to be very clear about the relevance and rank of different impacts dimensions and whether a large set of impact dimensions can really be achieved by one single measure respectively instrument. For policy design this means that the programme objectives have to correspondent with an appropriate mix of policy instruments and the right balance between direct and indirect funding. Additionally, policy design has to be very aware about the prerequisites for (behavioural and system) change which cannot entirely be influenced by singular measures.

1 Introduction

The question about the effects of policy interventions is a very basic one. It is of interest to politicians, who need to legitimize their work and who want to learn for the future; it is of interest to the public, who is interested in or addressed by the policy and whose taxes are used to pay for it; and finally, it is of interest to academics who are studying public policies.

Knowing about the effects of policy becomes more important the more money is spent for a particular measure. In recent years, this could be witnessed in the policy area of research, technology, development and innovation (RTDI) policies, where increasingly large amounts of money are being spent in order to move to what has been termed in the European context a "knowledge-based society". Since European leaders expressed this goal numerically in their 2000 Lisbon strategy, stating they wanted to increase the share of RTDI spending up to 3% of the Union's GDP, national governments have invented a number of new policy measures to foster this goal.

Legitimation of public funding is but one of the reasons for an increased demand for what is called impact assessments. Another reason seems to be that the knowledge about the effects of policies and learning from policies more generally have increased in value in order to improve future policy making. "The demand for impact assessment can be seen as one element of the move to a knowledge-based society. Policymakers need to know what the results of their past policies have been, and to have a better idea of what the probable results of future policies will be. The idea of impact assessment is an obvious response to this demand," (Miles, Cunningham et al. 2005: 141).

The assessment of impacts is an important function of policy evaluation, if not its central task. It requires the measurement of change and the attribution of the observed changes to the policy intervention, so, in a nutshell it needs to establish a link between the policy intervention and the observed effects. This has spawned intensive thinking in the evaluation community about how impact assessment should be performed, what concepts and methods appear to be adequate, and how to deal with certain theoretical and methodological challenges such as counter-factual reasoning, quantification or time-lags of effects. For evaluation in the field of RTDI policies, there are a number of compendium-like approaches which address these questions such as the SMART study (Miles, Cunningham et al. 2005), the RTD Evaluation Toolbox (Fahrenkrog et al. 2002), and others (Rhomberg et al. 2006, White 2009). In addition to the knowledge on how impact studies might be performed, the ImpLore project most recently (ImpLore 2009) gave insight to the question on what the impacts of innovation programmes are like: Do innovation programmes have a noticeable effect on innovation?

While these how-to-do-guides set out frameworks from an (applied) scientific viewpoint, one might ask how impact assessments are being performed in reality – the reality of service contracts for programme owners, most probably including budget restrictions, customer needs, and tough schedules. One might ask whether there is some systematic use of methods, i.e. whether there are certain sets of methods which are employed for specific policy measures and in specific contexts. One might also ask whether evaluation studies of policy programmes have an impact on future innovation policy.

This thematic paper is going to address these questions along four thematic lines:

- The Relationship of Evaluation and Impact Assessment

- The Performance of Impact Assessments
- The Impact and Usefulness of Impact Assessments
- General Developments and Challenges

1.1 Data Sources used

We use a mixed quantitative-qualitative approach, including first the quantitative exploration of the Inno Appraisal data set on variable relationships and general trends at a descriptive level, and secondly on a qualitative basis by conducting in-depth expert interviews in order to weigh our findings.

1.1.1 Descriptive Statistics

We analyze the Inno Appraisal database with respect to multiple variable relationships of impact assessments with certain characteristics of the evaluations. We use both samples of the dataset. Sample A consists of all templates received from programme managers as well as all pre-filled templates, if no template was received. Sample A with 171 cases is being used for all questions related to characteristics or topics and methods used in the evaluations (see sections 2 and 3 of this chapter), sample B with 132 cases represents the sets of templates validated by policy makers and is used for the analyses in chapter 4 on quality and usefulness (see chapter 2 of this report for more information on data collection).

1.1.2 Expert interviews

In this part we want to draw a line from the conceptual and methodological outset to the results and perception in the political arena. What is in particular useful for policy makers? Is it a high-quality impact assessment using specific approaches / methods or do other variables influence usefulness? For which purposes can impact studies be of use to policy makers?

Questions arising from / not answered in quantitative analysis were:

- What purposes are impact studies used for?
- Given the range of concepts and methods which are according to the data used in economic impact assessments, are there certain „basic“ requirements for impact assessments?
- What kind of impact study is the most useful one for policy makers?
- How important are impact studies for policy-making?

We ran a set of in-depth interviews with experts who have gained experience with a number of evaluation studies. In order to identify interview partners, we examined the data for certain sets of interesting evaluation studies. As the data have been collected country-wise and as the evaluated policy measures are of national scope, we identified country experts as interview partners who are able to oversee a number of appraisals collected in one country. Concretely, Interview partners were selected from countries with high number of good practice cases: Germany, Austria and UK. Overall, 10 out of 13 persons contacted (77%) responded to our request. The interviews took place between August and October 2009. The Interviewees (11 persons in 10 interviews) are from all three countries (6 German, 4 Austrian 1 from UK). Five of them are affiliated to ministries, three to project agencies and three are 3 evaluators / researchers. Their role in the policy cycle (multiple answers possible) was as follows: 9 have experience with the formulation of tenders, 7 with the selection of evaluators, 4 with Project management, and finally 8 are Evaluators. On average, each interviewee had experience with 11 evaluations. Experience with certain types of evaluations (multiple answers possible): 4 with Ex-ante, 5 with Interim, and 6 with Ex-post.

2 Relationship of Evaluation and Impact Assessment

2.1 (Operational) Definitions

In our work we focus on the impact assessment of RTDI policies which take the form of programmes, i.e. measures that mostly provide direct or indirect funding to certain target groups in order to support specific innovation activities.

In this study, we use the term impact assessment to describe in a general way all attempts in RTDI policy appraisals to cover the aspect of effects and the undertaking to attribute these effects to the programme under consideration.

Impact assessment is one form of research in evaluations (eg. White 2009) where analysis of added value is compared to the counter-factual situation (What would have happened anyway?) and in relation to the goals of the programme (Are the intended effects realized?). We understand impact assessment in the widest possible sense including all types of effects (Rhomberg et al. 2006: 12).

Typically, the effects of policy measures can occur at different points in time (short- mid- and long-term) and spread differently (i.e. at the direct level of participants or at the indirect level, i.e. beyond the participants). The effects can be distinguished among in the following terms:

- Outputs: : Short-term (Measurable) results of funded projects
- Outcomes: Effects on the participants of the programme
- Impacts: (Mid- or long-term) indirect effects, i.e. beyond the participants of a programme (spill-overs)

Again, we understand impact assessment in the widest possible sense. This includes all studies about the effects at the project level and at the level of the participants of a programme.

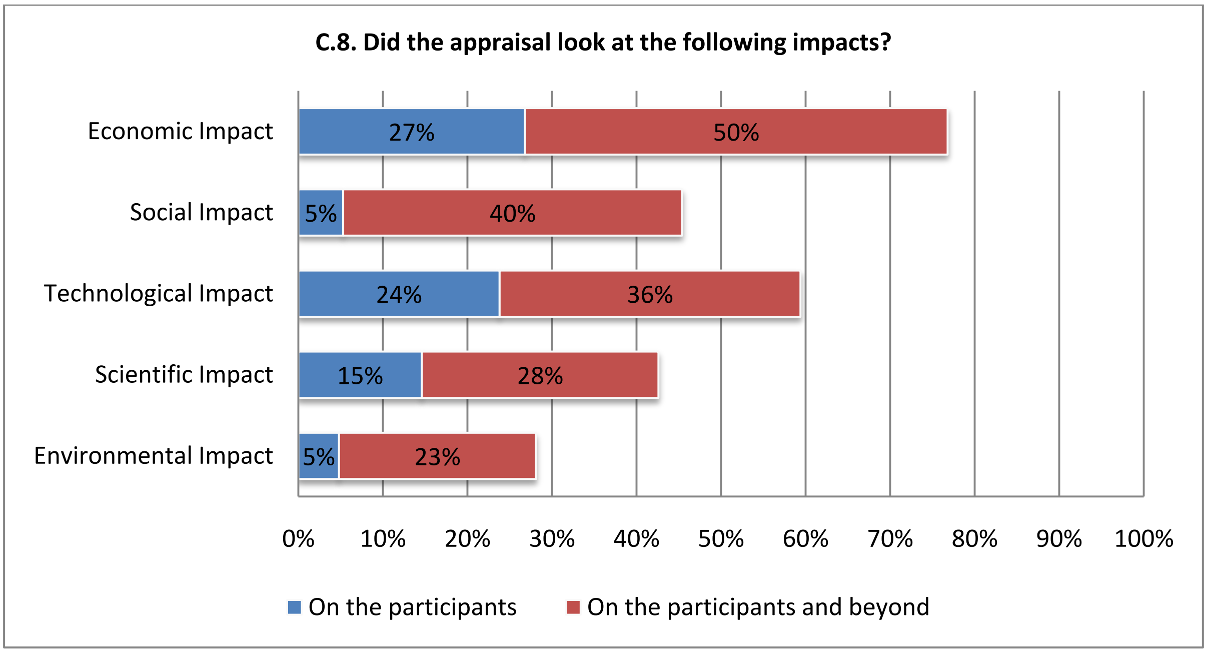
2.2 Importance of Impact Assessments for Evaluation

The INNO-Appraisal template covers five impact dimensions (economic, social, scientific, technological and environmental) and two types (direct and indirect impacts). Exhibit 40 shows the coverage of the different impact dimensions for the whole dataset (sample A). From Exhibit 40, we see that overall the most frequently covered impact dimension is the economic one: 77% of the evaluations cover economic impact. This is due to the fact that “classical” innovation policy and network programmes make up a large share of the dataset, which are designed to foster economic growth.

Almost half of the studies cover social impact, which can include many dimensions. It can refer to structural implications such as new jobs or the role of women in innovation processes. It can also refer to behavioural aspects such the establishment of sustainable networks, the promotion of innovation mentality, changes of risk attitudes, the awareness of societal needs, acceptance of technology or attitudes towards entrepreneurship. Almost every third evaluation covers environmental impacts.

In terms of numbers, technological impacts are also very important with 60% of evaluations covering this type of impact. 43% of evaluations cover additionally scientific impact. These numbers give the impression that the evaluations considered in our dataset are related in many cases to R&D programmes (as opposed to programmes which promote non R&D related aspects of innovation).

Exhibit 40: Impact Dimensions covered by the evaluations [69]



The data set displays two different types of impact assessment: direct (on the participants) and indirect (on the participants and beyond). Two aspects should be mentioned here:

- Indirect impact assessments are an aspect of quality: In evaluation research, there is the understanding that a sound impact analysis covers also indirect impacts. The ultimate aim of many programmes is to bring about sustainable changes in innovation performance – including spill-over effects – and this cannot be covered by focusing on direct effects only (e.g. SMART study 2005: 141). Half of the evaluations claim to cover indirect economic impact, 40% cover indirect social impact, which are rather high shares.
- The fact that only some few studies consider direct social or environmental effects might be explained by the consideration that social or environmental effects are by nature external, i.e. not restricted to the participants of a programme.

The two dimensions most frequently analyzed with respect to indirect effects are economic and societal impacts. Table 1 (in the Annex 6A) lists the appraisals, which cover economic and societal impacts per country. In 19 out of 22 countries, we can find indirect impact assessments, so we may conclude that this is a widely used understanding of impact assessment.

Overall, our expectation that the question of impact is central to evaluations can be confirmed: Table 74 shows that about 81% of the evaluations claim to cover impacts in a wider understanding.

---

[69] Percentage of responses to the mentioned categories in Sample A

On closer inspection, the economic dimension seems to be the most dominant, while other impact dimensions are covered additionally. In sum, only 7% of the evaluations cover impact assessments without covering economic impacts. Typically, evaluation studies of innovation programmes seem to cover more than one impact dimension.

Table 74: Economic impact coverage in relation to the number of impacts covered[70]

		number of impact dimensions covered						**Total**
		0	1	2	3	4	5	
	Evaluation covers economic impact		9%	19%	18%	16%	12%	74%
	Evaluation does not cover economic impact	19%	4%	3%				26%
Total		19%	13%	22%	18%	16%	12%	100%

2.3 Characteristics of Evaluations with Impact Assessments

Do certain types of evaluations use impact assessment? We are interested in finding out, whether certain conditions of the commission of an evaluation trigger specific types of impact assessment. When do we find economic, in which instances social impact? And are there conditions that require a large number of impact dimensions to be covered?

We are checking variable relationships with the most important characteristics describing the evaluations in the sample:

- number of measures evaluated: single measure vs. portfolio measure evaluation
- programme funding: national or structural fund
- tender procedure: internal, no tender, closed or open
- evaluator: internal or external
- timing: ex-ante, accompanying, interim, ex-post
- purpose: summative or formative
- external co-sponsorship
- planned evaluation (foreseen during the design phase of the measure)

These analyses are guided by the following assumptions:

a. Portfolio evaluations are regarding several policy measures at a time and aim at understanding their complementarities. Therefore we expect that their (combined) effects are considered in such evaluations. In particular, indirect effects (beyond the single measures) should be a topic.
b. Structural fund evaluations differ from the evaluations of national innovation programmes, as they are commissioned by the EU, i.e. externally. We expect them to cover a greater range of impact dimensions.
c. Competition in tender procedures could lead to the selection of a bid, which offers extra work, such as the coverage of more impact dimensions. Therefore, we expect a higher number of impact dimensions in open tenders.

---

[70] Number of appraisals in sample A, 31 missing values.

d. Impact assessment can and should rarely be based on mere numbers. It requires qualified assessments of evaluators – whose impartiality is higher, if they are external. Therefore we expect that impact assessment is rather commissioned to external than internal evaluators.
e. Although ex-ante impact assessment is an issue, the dominant nature of impact assessments is expected to be retrospective. In fact, the later the point in time when an evaluation takes place, the higher the probability that any effects are observable. In particular, indirect effects require often long time to unfold. We expect to find a timing bias in the data, i.e. a relationship of impact assessment with interim and ex-post evaluations.
f. Due to their character, summative Evaluations should cover a greater range of impact dimensions than formative evaluations.
g. Similar to SF evaluations, we assume that external co-sponsorship leads to a higher number of impact dimensions covered.
h. When an evaluation is planned during the design phase of a measure, we expect that impact assessment is a systematic element of the evaluation, while ad-hoc commissions of evaluations are rather restricted to certain aspects of evaluation – i.e. we assume that ad-hoc evaluations do not systematically include impact assessment.

Table 75: Characteristics of Evaluations and the coverage of impact[71]

		Economic Impact	**Social Impact**	**Number of Impact Dimensions covered**
Portfolio Evaluations				
SF Evaluations		Positive	positive	positive
Tender procedure	internal	Positive	positive	positive
	no tender	Negative	negative	negative
	closed	Negative	negative	negative
	open	Positive	positive	positive
Evaluator	external			
	internal			
Timing	ex-ante			
	accompanying			
	interim			
	ex-post			
Purpose	summative	Positive		negative
	formative	Negative		
	both			
External co-sponsorship			positive	
Planned Evaluation				

Table 75 gives an overview on the results of the analyses which are statistically significant. We find:

a. No support for a relationship of impact assessment and portfolio evaluations. The impact assessment in these evaluations does not appear to be remarkably different from single measure evaluations.

---

[71] All mentioned relationships are statistically significant (Pearson's chi square).

b. Strong support for a special importance of impact assessment in SF evaluations. SF evaluations are covering in many times economic and social impact. And structural fund evaluations cover a greater range of impact dimensions than other evaluations, often up to 4 or 5 dimensions, i.e. often, the environmental dimension is included.

c. There is a link of tender procedures to impact assessment:
    - Internal and open tenders tend to cover economic impact, in particular indirect; closed and no tenders very often do not cover economic impact.
    - Internal and open tenders also tend to cover social impact, while closed and no tenders do not.
    - Internal tenders most often cover 4 and 5 impact dimensions, closed and no tenders in most cases do not cover any impact dimension at all.

    The findings for open tenders are as expected. An explanation for the findings about internal tender procedures is that most internal tenders are used for SF evaluations. The results for closed tenders seem to be a country-specific phenomenon. Austria employs many closed tenders and at the same time we see that impact assessment is a rare topic in Austrian evaluations.

d. No significant difference between internal and external evaluators appears.

e. Surprisingly, also no timing bias can be witnessed. We can only explain this by the fact that ex-ante impact assessments seems to play a greater role than expected, in addition to the fact that interim and ex-post evaluations do not systematically cover impact.

f. Indeed, there is a difference between summative and formative evaluations, although slightly different than expected. Purely summative evaluations do cover economic impact (while formative more often do not), but the summative evaluations concentrate often only on one or two impact dimensions.

g. External co-sponsorship shows partly the expected pattern: if there is an external sponsor, the evaluation includes more often social impact assessment.

h. No special pattern for planned evaluations.

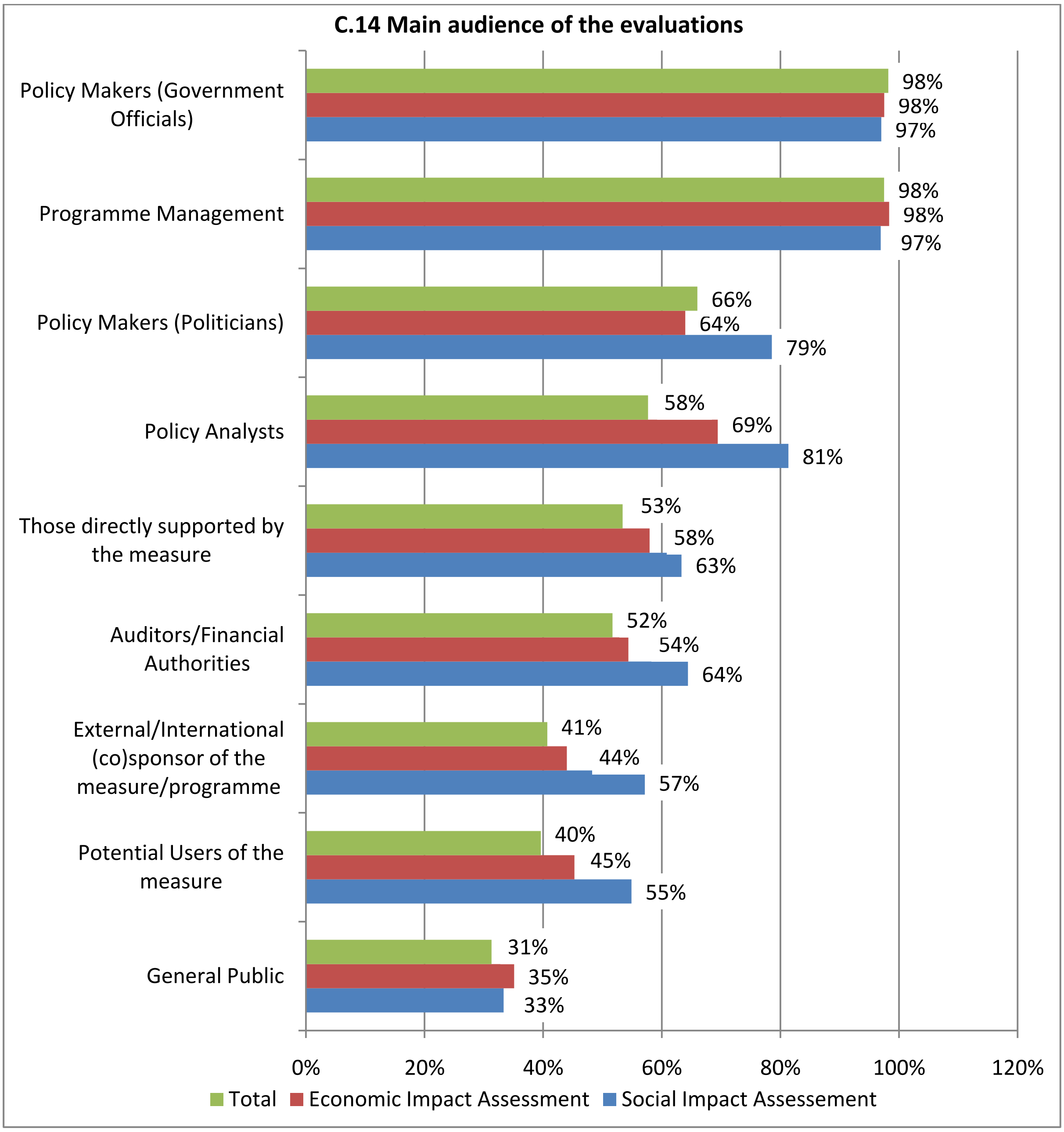
2.4 Audiences of Evaluations with Impact Assessments

Impact assessments address different audiences than evaluations on average do. Exhibit 41 shows the audiences of studies which cover economic impact assessment and studies which include social impact assessment compared to the audiences of all evaluations in the sample. Of course, impact assessments address like all evaluations the programme managers and policy makers at the working level.

Apart from this pattern, impact studies are on average more often used for external legitimation and communication. They are used for reporting to auditors (parliamentarians), to high-level policy makers and to external co-sponsors. The two latter results apply in particular to studies which cover social impact. Very often, policy analysts are mentioned as audience of impact studies. This is certainly the case, because there is a broad scholarly debate on methodological aspects of impact assessment.

Impact assessments also serve for communication to programme participants and potential users of the programme. To find overall higher audience rates for impact assessments points to the importance of the topic within an evaluation framework. As soon as there is some evidence about the effects of the policy measure, this is widely communicated. As programme owners should be inclined to communicate primarily positive aspects about their programme, we can conclude that our sample includes a lot of impact assessments which find evidence for impacts and arrive at a positive conclusion.

Exhibit 41: Addressed Audiences of Evaluations with Impact Assessments[72]



---

[72] **Reading Aid**: Exhibit shows Yes responses to the mentioned categories in Sample A. Data rows are sorted by the frequency the audiences are addressed in the total sample (in descending order), i.e. most evaluations address policy makers, while the general public is addressed only by a third. Reading Example for bars: 98% of the evaluations, which cover economic impact assessment, address policy makers (Government Officials).

2.5 The Use of Impact Assessments for certain types of policy measures

This section explores whether we find impact assessments for certain types of policy measures. It might be the case that not all programme types yield effects which can be the subject of an impact assessment. And while many programmes intend economic effects, there is not always the intention to reach at social effects.

We find that direct financial support for R&D, network and cluster programmes and programmes targeted at specific sectors and regions are quite regularly subject to economic impact assessment. This has partly to do with the fact, that these types are quite numerous in our sample, but it is at the same time a reasonable result. Impact assessment of regional programmes is obviously due to structural fund requirements.

Science-industry cooperation programmes are often subject to evaluations which include social impacts. This could be the case, because these measures intend to bring about new cooperation patterns and new forms of knowledge generation.

2.6 Results from expert interviews: Main aim of evaluation studies compared to main purpose of impact studies

We may compare the quantitative results with the indications given by some of the experts, guided by the overall question: What purposes are impact studies used for (compared to purposes of evaluation studies)? This is motivated by the often repeated fact that the demand for impact assessment has grown because of the need to legitimize political interventions.

Almost all experts agree that evaluations and impact studies have similar purposes: Both serve Learning and Legitimation. However, several experts point to the fact that impact studies are primarily used for legitimation purposes and less for learning, while the primary function of evaluation exercises is learning. Function is also dependent on type of evaluation as well as target group:

- In interim/ accompanying studies, learning is more important, in ex-post, legitimation prevails.
- For high-level policy makers, often legitimation of the policy intervention is most important, whereas for the operational level / project agencies, learning on how to improve the measure is more important.

It is important to mention that in reality most policy makers (and programme officers as well) do not differentiate between “impacts” as a crucial element within evaluation studies meaning that “effect” of a measure are described (typically through survey questions) and “impact assessments” as particular quasi-experimental design including control-group approaches etc. Therefore the message of the policy level is that every evaluation has to tell something about effects and “impacts” to meet both objectives – learning and legimitation - but that the legitimatory aspect dominates in those studies where the dimension of impact assessment is in the focus.

3 Performance of Impact Assessments

We have knowledge on how to do impact assessments from the literature and from practitioners in the field. Often, sophisticated methods (econometric modeling, control group approaches, quasi-experiments etc.) as brought forward by the literature are difficult to realize due to the lack of data or resources. Practitioners therefore show in their contributions some good practice examples for dealing with theses issues. On a broader picture, however, we do not know whether this knowledge is used by evaluators all over Europe.

To get an overall picture which topics and methods are linked to economic and social impact assessment, we analyze cross-tabulations of two sets of variables with economic and social impacts: topics covered in the appraisal as well as data collection and analysis methods employed.

3.1 Topics covered in Evaluations with Impact Assessments

Table 76 compares the application of topics in evaluations which cover impact assessments and in those which do not. We run two comparisons – for economic impact and for social impact assessment.

At first sight, we find a striking result: Evaluation studies, which cover the ***assessment of economic impacts***, employ almost the full variety of classical evaluation topics. For almost all topics, we find statistically significant[73] positive relationships. However, the interpretation is straightforward: Many impact assessments in our sample are embedded into broad evaluation approaches instead of focusing on impact alone. And, of course, many studies in our sample cover economic impact.

In more detail, this means that questions of design or concept of a policy measure are linked to economic impacts. Interestingly, External and Internal consistency as well as Coherence/Complementarity are primarily linked to impact studies which also cover indirect economic impacts. Although the issue of coherence is frequently addressed in economic impact studies, it is more often covered in studies which do not assess economic impacts. However, overall the results point to the fact, that many evaluations seem to take a very context-sensitive approach which analyses in different ways the context of the policy measure including its broader impact.

Management and implementation topics are also linked to economic impact assessments. We find Project Implementation Efficiency to be positively linked to impact assessments. Programme implementation efficiency on the other hand, is by far more often covered in studies without impact assessments. The first group again shows that we find broad evaluation studies in the sample, the second finding describes a group of evaluations which concentrate on implementation processes.

Classical impact assessment topics finally are of course linked to the coverage of economic impacts, such as Goal attainment/Effectiveness, Outputs/Outcomes/Impacts, Quality of outputs, Value for money etc, or Input, Output and Behavioral Additionality. Behavioral additionality is mainly linked to direct economic impacts, meaning that it is of interest to the evaluators to study how the behavior of participants changes. This goes in line with the fact that we find many interim evaluations in our sample. Because of the timing of an evaluation early in the lifetime of a project, it may not yet be possible to study input or output additionality, but there may be already effects to the behavior of participants.

---

[73] Based on Pearson's chi square.

Table 76: Topics covered in Evaluations compared to the coverage of impact[74]

Topics	Economic	Non-economic	Social	Non-Social
Outputs, Outcomes and Impacts	97%	76%	94%	89%
Goal Attainment/Effectiveness	95%	74%	94%	85%
Internal Consistency	82%	74%	87%	73%
External Consistency	80%	71%	84%	71%
Policy/Strategy Development	77%	74%	86%	68%
Programme Implementation Efficiency	71%	89%	78%	71%
Quality of Outputs	68%	25%	66%	47%
Input Additionality	63%	13%	53%	43%
Output Additionality	61%	18%	56%	43%
Coherence/Complementarity	60%	68%	69%	54%
Project Implementation Efficiency	57%	19%	55%	38%
Behavioural Additionality	55%	37%	47%	48%
Value for Money/Return on Investment/Cost-Benefit Efficiency	37%	3%	36%	19%
Gender issues	25%	16%	37%	9%
Minority issues	9%	3%	14%	1%

The set of topics linked to ***social impacts*** is quite different from these results. Social impacts are also studied in evaluations which cover context topics (external consistency, coherence/complementarity). In social impact studies, coherence/complementarity plays a greater role than in studies which do not study social impacts. The link of social impact to studies which take a broader picture is also underlined by the fact that the topic of policy /strategy development is positively related to the coverage of social impacts. Apart from that, we find a link to the quality of outputs while other classical impact assessment topics do not show any significant relationships – except for the quantitative concept of value of money. We assume that the use of this concept is not caused by the aim to study social impacts, but rather by the fact that studies which use this concept to study economic impacts, cover at the same social impacts.

Finally, gender and minority issues play a role for social impact assessment, which is most likely due to the requirements of structural fund evaluations. Outside SF, we do rarely see the coverage of gender or minority issues.

3.2 Methods used in Impact Assessments

Economic and social impact assessments use more or less similar methods. While we have found for evaluations which cover economic impacts without covering social impacts a different set of topics employed than for those evaluations which cover both impact types (see 3.1), we find in this section that these two groups of evaluations do not so much differ in terms of methods applied.

For data collection methods, this means more precisely, that the most classical methods – participant surveys and interviews – are linked to economic and social impacts. When evaluations cover social impact, there is also a link to the use of existing surveys. We explain with the character of social impacts, which is in most cases indirect, i.e. beyond the participants of a programme. In order to cover aspects outside the programme, often the use of additional (existing) data sources is necessary. The use of technometrics/bilbiometrics is also linked to social impacts, but this finding should be interpreted with care, as there are too few cases which employ these data at all.

---

[74] Shaded cells show significant relationships between column and row variables.

Table 77 shows the results of cross-tabulations of impact types with data analysis methods. We find a large set of data analysis methods linked to the study of impact. More or less all of them would be straightforwardly expected to be linked to impact assessment.

Evaluations which study ***economic impacts*** use econometric modeling, Input/output analysis, network analysis, before/after group comparisons. As main design approaches, they use either cost-benefit approaches, counter-factual designs or control group approaches. The result for some of the methodologically advanced practices (control groups, before/after group comparison) should however be qualified by the aspect that the number of studies which employ these methods is quite small.

We find also that case studies and context analysis are linked to economic impacts, which tells us that economic impact assessment is either performed also at a qualitative level or within evaluations that consider the broader context.

Table 77: Methods covered in Evaluations compared to the coverage of impact[75]

Data	Economic	Non-economic	Social	Non-Social
Descriptive Statistics	79%	70%	77%	74%
Context Analysis	70%	55%	77%	59%
Document Analysis	52%	45%	47%	52%
Case Study Analysis	45%	26%	51%	30%
Input/Output Analysis	34%	3%	37%	13%
Cost/Benefit Approach	31%	0%	40%	9%
Econometric Analysis	29%	3%	24%	17%
Counter-Factual Approach	26%	11%	20%	20%
Control Group Approach	25%	3%	19%	16%
Network Analysis	20%	8%	25%	9%
Before/After Group Comparison Approach	13%	3%	19%	2%

Evaluations which cover (in addition) ***social impacts*** use primarily a subset of these methods. Social impact is linked to case studies, context analysis, network analysis, input/output analysis, cost-benefit approach and before-after group comparison. To find case studies and network analysis (and also before/after group comparison) linked to the study of social impacts seems straightforward, as social effects often refer to behavioral aspects which can be covered by these methods. The link to context analysis underlines again that many evaluations are paying attention to context, and this seems to be particularly helpful for the assessment of social impact. To find particular quantitative methods linked to social impact study can be explained by the fact that social impact assessment almost always goes in line with economic impact assessment. Studies which show these particular methods for economic impact assessment and which cover at the same time social impacts, can be found among SF evaluations as well as in the UK and Sweden.

---

[75] Shaded cells show significant relationships between column and row variables.

3.3 Results from expert interviews: Conceptual and methodological requirements for economic impact studies

The interviewed experts fully agree that a Multi-Method Approach works best in evaluation studies general and impact assessments in particular. Additionally they mentioned that one can differentiate between two types of studies:

- Mainly Qualitative: here typically causal relationships and the counter-factual situation are dsicussed and non-quantifiable effects analysed; these approaches are often part of „classical“ evaluation studies
- Mainly Quantitative where control group approaches and quantifiable effects are in the centre of analysis and which typically represent separate studies. Generally this type if study is rarely used, because of data availability problems.

This view seems plausible, and it meets partly the findings from the data. According to the percentages of evaluations which cover certain evaluation topics and use certain methods, it seems that evaluation studies, which have a broad approach (e.g. context analysis) are the dominant type. However, these studies do not seem to be exclusively linked to qualitative methods, as we see that quantitative methods are used quite frequently, too. So, it seems, that evaluation studies often apply a multi-method approach, combining qualitative and quantitative data, while there are no big surprises about the types of methods used. For example, the use of standard methods such as participant surveys and interviews has been found in the data and was also mentioned by the experts.

4 The impact and usefulness of impact assessments

4.1 Evidence from the evaluation database

With the rise of the idea of evidence-based policy making (e.g. Nutley et al. 2002, Solesbury 2001, Sanderson 2002) in the UK, expectations have grown on the use of scientific evidence in policy making. Sound impact analyses might be valuable not only in a retrospective way, but also for policy change. As impact assessments have come up because of an increasing demand for robust (quantifiable) statements on the effects of policy interventions and in the context of evidence-based policy making, the expectations towards the studies might be very high. At the same time, establishing causal relationships between policy interventions and observed changes poses a theoretical challenge as well as empirical / methodological problems with the specification of the control group or counter-factual situation.

Because of this trade-off we might expect that policy makers

a. are not very satisfied with the application of the methods.
b. are not very satisfied with the performed analyses.

At the same time, policy makers might regard impact studies as a useful learning tool. Because of the retrospective nature of impact assessment, learning for the evaluated measure might be limited, but the evaluations could be of use for the design of other programmes or general policy formulation. Drawing more general conclusions from evaluation might work best, when the evaluation results are presented well placed into the context of the evaluated policy measure. Impact analysis might contribute to such a context analysis, therefore we expect that policy makers

c. are satisfied with the coverage of broader context.
d. consider impact assessments to be less useful for the evaluated programme itself.
e. consider impact assessments to be useful for the design of other policy measures or general policy-making.

We asked policy makers and programme managers to assess the quality and usefulness of the evaluation reports. We differentiated between nine aspects of quality related to an evaluation study and five dimensions of utility. For these questions the policy makers could give estimates on a five-point Likert scale ranging from "not at all" (value 1) to "definitely/extensive" (value 5).

Table 78 shows first the results for the quality aspects. We find that if impact assessment is included in evaluation, this leads in most cases to higher quality scores. In some cases, these scores are remarkably higher. Aspects of quality for which this difference turns out to be statistically significant are "addressing the terms of reference", "application of quantitative methods", "coverage of broader context" and finding the "conclusions based on the analysis". These results refer to those cases which cover economic impacts. Studies which cover social impacts appear also to be of a high quality, although they contribute significantly to only one aspect of quality – the "coverage of broader context".

Table 78: Quality of Evaluations with Impact Assessment[76]

	Mean economic	**Mean non-economic**	**Mean social**	**Mean non-social**
address TOR	4,39	3,85	4,35	4,10
design appropriate given the objectives	4,15	3,93	4,14	3,97
methods satisfy the TOR/purpose	4,29	4,30	4,27	4,27
application of qualitative methods	3,99	3,83	4,00	3,83
application of quantitative methods	4,05	3,43	3,86	3,83
information sources well documented	4,07	4,13	4,12	4,08
analysis based on given data	4,34	4,20	4,27	4,26
cover broader context	3,65	3,10	3,75	3,18
conclusions based on analysis	4,48	4,17	4,31	4,34

---

[76] Shaded cells show significant results from T-test.

Before discussing these results in view of the formulated expectations, we sum up the results for the usefulness of impact assessments.

There is no clear link between impact studies and the different aspects of usefulness in the dataset. This means: Studies that cover impact are not regarded as more useful than studies that do not cover impact.

We find higher usefulness of economic impact assessments for the design of contemporaneous and future programmes, but this is not above any level of statistical significance. Moreover we find studies which include social impact assessment appear to be less useful for the management and implementation of the evaluated measure.

Looking at the consequences, there is a quite clearer picture. In short, impact assessments are linked to certain consequences:

- Economic impact assessments are associated with the expansion/prolongation of measures.
- The assessment of indirect economic and social impacts is linked to "re-design of another measure".

Linking all these findings to the expectations, we arrive at the following conclusions for the impact and usefulness of impact assessment:

a. Against our initial expectation, policy makers are satisfied with the application of quantitative methods in economic impact assessments.
b. Policy makers value highly the conclusions from analyses of evaluations which include economic impact assessment. Moreover, policy makers think that economic impact assessments are addressing very well the terms of reference. We interpret from this, that policy makers seem to know what to expect from impact assessment.
c. Clearly, impact assessment is linked to high quality scores for the "coverage of broader context". This is true for economic and social impact assessment. We have learned in previous sections, that impact assessement is quite often carried out in combination with context analysis. Our expectation seems to be true that this contextual embedding of impact analyses contributes to higher quality of the evaluations.
d. Indeed, impact assessments are less useful for changes to the evaluated programme itself (internal usefulness). We assume that this is due to the retrospective nature of impact assessment. However, we find that impact assessments are linked to consequences of evaluations, in particular to the expansion or prolongation of a programme. So, the evidence of impact of a policy measure is a crucial information for the decision on extension/prolongation.
e. External usefulness (for other policy measures or broader policy formulation) was expected for impact assessments, and there is some supporting evidence. Although not statistically significant, we find higher usefulness scores of economic impact studies for the design of contemporaneous and future programmes. We also find – at a statistically significant level – that the assessment of indirect economic and social impacts is linked to "re-design of another measure".

4.2 Results from expert interviews

4.2.1 What kind of impact study is the most useful one for policy makers?

Generally spoken, it is important to take the Timing or „Window of opportunity“ into account. From the point of view of the interviewed policymakers and project officers, formative interim / accompanying studies have a higher chance of good timing. Also crucial - besides a sound empirical analysis based on at least some sort of quantitative data - is the continuous discussion between commissioners, evaluators, and stakeholders as well as a ongoing presentation of milestones and interim results of evaluation.

Further success factors for an improved usefulness of impact analysis are the identification of respectively a focus on central goals of the measure (What can be really an effect of R&D funding?) and, particularly for the policy level and the upper hierarchies simple and short statements, supported by quantifiable results.

Finally, the involvement of the stakeholders of a programme (at least selected representatives) might lead to increased usefulness.

For the operational level and implementation, a sound analysis is important, including broad documentation. Additionally, recommendations need to be tailored to what is in the power of the policy makers, i.e. the instruments/ design-elements of a programme.

These aspects mentioned by the interviewees go beyond what we can find in the data. The usefulness of an evaluation is clearly not (only) dependent on the application of methods or quality of analysis, but on the political context. This is not surprising. However, we have seen that the evaluations collected in the INNO-Appraisal data base already seem to respond to the need for quantifiable results of impact analyses. Often, quantitative methods are employed, and studies which cover economic impact analyses receive higher quality scores for quantitative analysis.

4.2.2 Influence of impact studies in decision-making processes

The interview partners were rather hesitant with their general assessment of consequences. If at all, the direct consequences of an impact assessment are minor and refer to the re-design and / or prolongation of a measure. However, their general view is that evaluation as part of efforts towards more transparency in policy-making can help to improve policy formulation. Policy makers have a motivation to perform well, for this reason policy learning takes place.

Of course, policy makers have to follow other rationales, too. Most important is the interest accomodation process, where evaluation results can serve as arguments. In particular, impact studies can legitimize budget decisions.

Again, these findings from the experiences of the interviewees add to what we can find at the case level of evaluations in the database. However, findings from the data and the expert interviews go in line as far as the prolongation of measures are concerned. We found, that impact assessment can contribute to the decision to expand or prolong a programme.

5 Conclusions

We have learned from this chapter a number of interesting facts. Focussed on the quantitative analysis of the data as well as expert interviews we found that:

General results

- Impact assessment is a central function of evaluation studies: A large number of studies across Europe claim to do impact assessment, currently most important are economic impacts.
- Impact assessments appear to be **central and wide-spread** across Europe
- Impact studies of structural fund evaluations differ significantly from impact studies of national innovation programmes.

Impact types

- Typically we find the use of a very **broad definition** of impact assessment, including all types of effects
- Assessment of economic impact is most dominant, other impact types of importance are technological and societal impacts (not: scientific and environmental impacts)
- The assessment of **new impact types** (apart from economic or technological) is still rather seldom. Societal impacts are often covered with an estimation of new jobs having been created, but other topics, such as gender impacts are quite rare.
- A high number of evaluations claims to assess **indirect impacts**, i.e. spill-over effects beyond the participants of a programme. This is given the methodological difficulties for assessing economic or societal impacts a surprising result. This result seems to reflect the demand for results on these spill-over effects.
- **Additionality** concepts are well established beyond the UK. They are employed by half of the evaluations in the sample. This is also true for behavioral additionality which has obviously become an integral part of the idea of additionality.
- Structural fund evaluations more often cover social and environmental impacts.

Methods used

- Almost the whole toolbox of possible **methods** is employed for impact assessment, including elaborate methods such as a control group approach.
- Most of the impact assessments are qualitative and part of broader evaluation studies.
- There are only few quantitative impact assessments using elaborated quasi-experimental designs like control-group approaches.
- Impact assessment is typically not a mere econometric exercise, but **often used in a** contextually **sensitive way.**

Policy Cycle

- Impact assessment is not a clearly retrospective element of evaluation. Often, it is also used in the form of ex-ante impact assessment and in accompanying evaluations.

- Evaluations which include impact assessments, in particular the assessment of societal impacts, are more often used for external communication. Experts confirm that impact assessment is in particular important for legitimizing the political interventions.
- If impact assessments are included into evaluations this leads to higher quality scores.
- With respect to usefulness, evaluations of (single) national programmes seem to be more useful for policy makers than structural fund evaluations.

The analysis and the interviews indicate a set of recommendations. Most important issues from policy maker perspective are:

1) Impact assessments are an important part of evaluations, but often not the only one. It seems that impact assessments seem to make most sense within the framework of a "normal" evaluation study which for example covers context analysis extensively.
2) Evaluators have responded to the demand for quantitative results and employ a variety of (elaborate) methods to achieve them. However, in most cases it seems that the combination of qualitative and quantitative analysis can cope more adequately with impact assessments, as many impacts are not quantifiable at all.
3) Many pitfalls of impact studies can be avoided by a constant communication between policy makers and evaluators during the process of evaluation. This leads to transparency for the whole evaluation process in order to realize learning and to cope with methodological challenges.
4) As impact assessments clearly pursue the two purposes of learning and legitimation, two types of recommendations might be considered: Those designed for policy improvement implemented by the programme owners / managers and those directed to higher levels, which serve the legitimation aspect.

For the future, we think that it is useful to consider further impact dimensions to a greater extent than in the past. For example, economic Impact assessment is often intended, but not possible due to long-term effects and complex environments. Additionally we expect more mission oriented policy programmes where other topics like sustainability, customer needs and the structural / regional development might become more important. Looking at the demographic challenges in most European countries, the issue to integrate larger parts of society to the research sector will become even more relevant than in the past and therefore impact assessments should address gender and minority issues as well in more detail. Finally, the still prominent aspect of Behavioural Additionality in Innovation Programmes (e.g. innovation management, risk aversion) will remain important.

For impact assessment, this all means that it will become even more demanding to measure the intended effects – at least quantitatively. Given that non-economic impacts will gain more and more in importance this would mean that new sets of criteria and indicators will have to be defined, and most likely many of these indicators will be of a qualitative nature. So, quantitative method development is – although academically interesting – not the only promising way to go but again a method-mix between quantitative and qualitative approaches. More public support for experimental evaluation designs (including meta-evaluations at national as well as European level) could help to identify the most promising ways to identify new impact types.

However, impact assessment will even more than today require that from the complex set of programme goals one has to be very clear about the relevance and rank of different impacts dimensions and whether a large set of impact dimensions can really be achieved by one single measure respectively instrument. For policy design this means that the programme objectives have to correspondent with an appropriate mix of policy instruments and the right balance between direct and indirect funding. Additionally, policy design has to be very aware about the prerequisites for (behavioural and system) change which cannot entirely be influenced by singular measures.

References

ImpLore (2009). “Benchmarking Strategies and Methodologies of National, European and International R&D Programmes, to Assess and Increase their Impact on Innovation”, Report to Lot 2 of European Commission Tender ENTR/04/96. April 2009.

Miles, Ian, Paul Cunningham et al. (2005). “SMART Innovation: A Practical Guide to Evaluating Innovation Programmes”, A study for DG Enterprise and Industry. October 2005.

Nutley, Sandra, Huw Davies und Isabel Walter (2002). “Evidence Based Policy and Practice: Cross Sector Lessons From the UK”. Hrsg. ESRC UK Centre for Evidence Based Policy and Practice. Working Paper 9. London. http://www.evidencenetwork.org

Fahrenkrog, Gustavo, Wolfgang Polt, Jaime Rojo, Alexander Tübke, Klaus Zinöcker (2002). “RTD Evaluation Toolbox- Assessing the Socio-Economic Impact of RTD Policies”, European Commission.

Rhomberg, Wolfram, Claudia Steindl and Matthias Weber (2006). “Neue Entwicklungen im Bereich der Wirkungsanalyse und –abschätzung FTI-politischer Maßnahmen“, Endbericht, ARC-sys 0108, Dezember 2006.

Sanderson, Ian (2002).. “Evaluation, Policy Learning, and Evidence Based Policy Making“. Public Administration (80) 1: 1–22.

Solesbury, William (2001). “Evidence Based Policy: Whence it Came and Where it’s Going”. Hrsg. ESRC UK Centre for Evidence Based Policy and Practice. Working Paper 1. London. http://www.evidencenetwork.org

White, Geoff (2009). “Pushing the boundaries of impact evaluation. Report on knowledge development possibilities”, SQW Consulting. April 2009.

PRO INNO EUROPE
INNO
APPRAISAL

Part II
Chapter 6

Exploring the Use of Behavioural Additionality

Behavioural additionality is a concept that is more and more common in evaluation of innovation policies, yet still often ill-conceived. This section analyses the application of behavioural additionality. It first gives a conceptual introduction based on the vast theoretical and empirical literature, describes how the concept is used and how evaluations using it differ from other evaluations in the INNO-Appraisal database, goes on to show that evaluators still struggle to define and operationalise the concept and finally gives some indication about how the concept can be used to greater effect.

Abdullah Gök and Jakob Edler[77]

---

[77] The lead authors for this section of the report are as given above, with contributions from all other team members at different parts of its production, conclusions and interpretations have been approved and are shared by the group.

Table of Contents

Table of Contents 152
List of Tables 153
Table of Exhibits 153
Executive Summary 154
1 Introduction 157
2 The Concept of Behavioural Additionality 157
3 Statistical Analysis: What differs in evaluations with behavioural additionality 162
  3.1 Introduction 162
  3.2 Coverage of Additionality 163
  3.3 Measure Types and Target Groups 165
  3.4 Country Distribution 167
  3.5 Structural Funds and Portfolio Evaluations 169
  3.6 Evaluator and Tender Process 169
  3.7 Timing and Purpose 171
  3.8 Budget and Planning of Evaluations 172
  3.9 Impacts Looked at in Evaluations 174
  3.10 Sponsors of Evaluation 175
  3.11 Evaluation Designs 176
  3.12 Data Collection Methods 177
  3.13 Data Analysis Methods 179
  3.14 Main Intended Audiences of Evaluation 180
  3.15 Terms of Reference 181
  3.16 Quality of Evaluations 182
  3.17 Usefulness of recommendations 185
  3.18 Consequences of Evaluations 185
4 Text Analysis 187
5 Case Studies 191
  5.1 Introduction 191
  5.2 The cases 192
6 Conclusion 197
References 199

List of Tables

Table 79: Classification of Behavioural Additionality Definitions in the Academic Literature 161
Table 80: Summary of the Perceived Quality of Additionality.. 183
Table 81: Classification of Definitions of Behavioural Additionality in the Practice 190
Table 82: The Cases of behavioural additionality 192

Table of Exhibits

Exhibit 42: The Evolution of the Concept of Behavioural Additionality 160
Exhibit 43: Uptake of Additionality as a Topic of Evaluation 163
Exhibit 44: Euler Diagram of the Coverage of Types of Additionality in Combination 164
Exhibit 45: Dendrogram of Topics Covered (Average Linkage Between Groups).............................. 165
Exhibit 46: Measure Types and Additionality 166
Exhibit 47: Measure Target Groups and Additionality... 167
Exhibit 48: Country Distribution of Evaluations and Additionality 168
Exhibit 49: Structural Fund and Portfolio Evaluations and Additionality 169
Exhibit 50: Type of Evaluator and Additionality... 170
Exhibit 51: Tender Process of Evaluation and Additionality 171
Exhibit 52: Timing of Evaluation and Additionality 172
Exhibit 53: Purpose of Evaluation and Additionality 172
Exhibit 54: Condition of External Sponsorship and Additionality 173
Exhibit 55: Dedicated Evaluation Budget and Additionality 173
Exhibit 56: Foreseen and Planned Evaluation and Additionality 174
Exhibit 57: Impacts Looked at in Evaluations and Additionality 174
Exhibit 58: Sponsors of Evaluation.. 176
Exhibit 59: Evaluation Designs 177
Exhibit 60: Data Sources and Collection Methods and Additionality 178
Exhibit 61: Data Analysis Methods and Additionality 180
Exhibit 62: Main Intended Audiences of Evaluation 181
Exhibit 63: Terms of References of Evaluations... 182
Exhibit 64: Quality of Evaluations and Additionality.. 184
Exhibit 65: Usefulness of Recommendations of Evaluations and Additionality 185
Exhibit 66: Discussion of the Results and Additionality 186
Exhibit 67: Consequences of Evaluations and Additionality 187

Executive Summary

Behavioural Additionality is still a rather novel, but already key topic for evaluations. The concept has enlarged our thinking about the effects of innovation policy to include more systematic learning as a key outcome in itself, enabling further, and broader, and more sustainable innovation. Behavioural evaluation seems to be a case of reflexive intelligence, whereby working on understanding the concept and applying it to innovation policy itself co-evolves with innovation policy concepts that take learning into account much more profoundly. Evaluation practice and conceptualisation on the one hand, and innovation policy development on the other hand have re-enforced each other. The empirical analysis in this section rests on three pillars, a statistical analysis, a text analysis of evaluation reports, and a set of interview based case studies of evaluations.

As a starting point, the section summarises key literature on behavioural additionality concept and history. In this literature, the term is understood in at least four different ways including i) an extension of input additionality, ii) the change in the non-persistent behaviour related to R&D and innovation activities, iii) the change in the persistent behaviour related to R&D and innovation activities, and iv) the change in the general conduct of the firm with substantial reference to the building blocks of behaviour.

Against this background, a text analysis of 33 selected evaluations demonstrated that in evaluation practice there are also at least four, more or less distinct, understandings of the concept in evaluation practice. Thus there is yet, no dominant understanding established just as the case in the scholarly literature. These four types are not explicit in most of the evaluation reports, as the concept is often not defined clearly but with small adjustment they correspond to the division in scholarly literature. For those cases in which there is no explicit definition, the types can be constructed form the practices, the approaches, and intentions mentioned in the report. The types overlap, but not entirely match four ideal types as defined in the vast literature on the concept. They are distributed rather evenly through the sub-sample and differ in their conceptual outreach, ranging from collaboration (non-persistent) in R&D and innovation only – the most narrow type – to persistent change in management practices more broadly, beyond R&D and innovation.

The analysis of the INNO-Appraisal database aims to show if and how evaluations differ that apply the concept from those that do not. For the first time this allows us to acquire a systematic picture of the nature of behavioural additionality in practice. The core results are as follows:

The data analysis shows that behavioural additionality is a well established concept in evaluations: 50% of all reports in the database employ it, explicitly or implicitly. The concept is more often used for networking and technology transfer concepts, which is consistent with the need for learning, networking and cooperation in those programmes. The behavioural additionality concept is most often used in conjunction with input and/or output additionality. It appears to be more important in evaluations that are also concerned at project level evaluations, not only at programme level, which again is consistent with the basic idea of understanding the micro level in order to understand the macro effect. The concept is less common in portfolio and structural fund evaluations as those often do not look at the project level.

While there is no difference between evaluations that are sponsored by the programme owners themselves or by other bodies, we observe that the concept is slightly less often applied in internal evaluations. The application needs specific expertise and in-depth qualitative approaches, which seem to be best conducted by external evaluators. However, this does not imply that evaluators are more keen to apply it than policy makers, since the concept is more often applied in those evaluations in the database that specify the methodology in the terms of reference and thus express a clear demand for behavioural additionality approaches. Our in-depth case studies indeed confirm that both evaluators and policy makers can be the source for the application of the concept, it is not entirely evaluator driven.

Interestingly, and neglecting its full potential, behavioural additionality is not as common in accompanying evaluations as one would assume given the focus on interaction and learning and the need to re-adjust programme and implementation should learning effects not be observed in real time. The concept is used in formative evaluations, but not as extensively as one would think. Similarly, evaluations that cover behavioural additionality are less likely to look at social and environmental impact, but much more at scientific and technological impact than the whole sample, while there are no differences as for economic impact.

As for methods, behavioural additionality evaluations are more qualitative and apply those methods with greater quality, however, the extent of case study analysis is not as broad as one would expect. Behavioural additionality evaluations also use surveys more often, while they cannot rely on existing data or monitoring data, pointing towards a need for adjusted monitoring.

Behavioural additionality evaluations are broader discussed across government and beyond government, and they are more often targeted towards the general public and towards users. All this points to the learning and mobilisation potential of the concept. However, evaluations applying behavioural additionality are not perceived to be significantly more useful for changes in policies than other evaluations (although they perform slightly better). In terms of concrete consequences of the evaluations that apply behavioural additionality, the major difference to the general dataset is that the former lead significantly more often to the extension of existing measures. This again points to some understanding of long term effects and the need for time in programmes that rely on the learning of actors.

The case studies finally confirm the variety of understandings, and different applications of the concept, and the challenges the application of the concept faces. This is true both at the receiving end, the programme owners, and at the performing end, the analysts. The cases show that evaluators and policy makers alike are keen to understand changes in behaviour better, but they also confirm that policy makers strongly demand a demonstration as to how the behavioural change translates into the intended innovation effect. However, many variables influence change in innovation and attribution remains a constant challenge as innovation effects often take considerable time to realise. Evaluations thus must clearly demonstrate the conceptual link between behavioural change and the innovation effect. They then must empirically grasp the change in behaviour and try to find robust indications that the conceptual link to innovation effects exists.

As yet, the applied methodologies do most often not fully grasp behavioural additionality. The cases however show that it is possible to differentiate behavioural additionality and define building blocks of behaviour as well as a chain of effects. This can be done in a mix of deductive and inductive approaches, with a focus on interaction with the beneficiaries. But there is also a delicate balance between exploring the concept to its full potential through all sorts of differentiation and methodologies on the one hand and pragmatic considerations and limits of absorptive capacity on the other hand. Thus, more experiments with sophisticated methodologies are called for. Those experiments should then enable us to define sets of meaningful, simplified methodologies that are more effective and efficient than the existing approaches, but do not overburden the process. To that end, there seems to be a huge potential in improving monitoring of programmes to use it for evaluations much more thoroughly.

Finally, the complexity of behavioural additionality asks for a strong interaction and communication between those commissioning the evaluation and the evaluators, since key concepts as to the link of behaviour changes to innovation must be shared between them and expectations clarified early on. Sophisticated methods alone do not guarantee the full benefits of the concept, their applications and the results must be intensively discussed among all stakeholders involved.

1 Introduction

Behavioural additionality is widely defined as the persistent change in the behaviour of the agents (firms in the case of innovation policy) which could be exclusively attributable to the policy action, i.e. the behavioural change that could not have happened had they not been supported. After its coinage in 1995 (Buisseret et al., 1995), the term has gained considerable attention first in scholarly literature and subsequently in evaluation practice in the domain of innovation policy. The OECD project in which a number of member states conducted pilot studies to evaluate behavioural additionality in their programmes marked another phase for the concept.

In spite of its increasing uptake, the term is not yet maturated as the definition and theoretisation of the concept still needs further work (Gok, 2010). The literature presents wildly different and sometimes conflicting perspectives on the concept and these results in further difficulties in its evaluation. Therefore, the very question about the concept is still open: what exactly is behavioural additionality. This case study ultimately intends to shed some light on this by first studying the distinct characteristics of behavioural additionality, i.e. the traits that make behavioural additionality evaluations different than other evaluations. This will be done by a statistical analysis of the INNO-Appraisal database. Furthermore, the case study will discuss a text analysis of the evaluations that use the concept to understand how exactly the term is used. Finally, five case studies of evaluations that covered the concept of behavioural additionality will be discussed not only to pursue the above question but also to gain insight into the use and usefulness of the concept for operational learning and policy feedback by contextualising it.

2 The Concept of Behavioural Additionality

The concept of behavioural additionality, which was originally coined in 1995, (Buisseret et al., 1995) has been subject to a certain number of scholarly publications. As it can be seen from the evolution of this literature shown in Exhibit 42, between 1995 and 2006 there were a number of papers which mainly dealt with the original idea. The OECD study followed by a book in which 11 case studies for different member states marked the beginning of the operationalisation phase of the concept. Finally, the recent attempts benefit from these conceptual and operational advances.

As well as being a concept for evaluation of innovation policies and programmes, behavioural additionality is at the heart of the policy discussions. The concepts of input and output additionality are widely considered as the hallmark of the neoclassical policy rationale which ultimately seeks to remedy the market failure by re-instating the second-best. Therefore, any policy with neoclassical rationale should create input and/or output additionality. On the other hand, behavioural additionality is considered as the core of the evolutionary/structuralist view which urges policy action to overcome system failures. In this view, a policy action is successful only if it changes the persistent behaviour of the agents, i.e. creates behavioural additionality (Bach and Matt, 2002, 2005; Lipsey, 2002; Lipsey and Carlaw, 1998a, 1998b, 2002; Lipsey et al., 2005).

Although the concept has gathered considerable attention from a range of scholars, there is still no consensus as to what it means. Similarly, the concept still severely lacks a theoretical basis. Gok (2010) classifies these definitions into the following categories which are also summarised in Table 79:

A. Behavioural Additionality as an Extension of Input Additionality:

There are a number of papers that understand behavioural additionality as a very simple concept that complements the excessively linear and strict nature of input additionality. For instance, Luukkonen (2000: 713) argues “input additionality and behavioural additionality are usually merged together in a question that lists different degrees of additionality, whether the R&D would not have been carried out at all without public support, or alternatively whether the public funding changed the scale and scope of the R&D or R&D would have been done differently”. Similarly Hsu and his colleague’s use the very same definition in their empirical articles (Hsu et al., 2009; Hsu and Hsueh, 2009).

Other scholars accept that there might be further effect,s although they either put the emphasis on the extensions of input additionality or they find only this one workable (evaluatable). Falk (2007), who defines behavioural additionality as a broad category that includes scope and acceleration additionality as well as cognitive capacity additionality and uses only the first two in his empirical investigation, is an example of this. Another example is Malik et al. (2006: 206). They accept that behavioural additionality is a multi-layered concept; all the same, they use and prefer the simplistic definition. Finally, Georghiou (2002a: 59) defines behavioural additionality as the superset of scale, scope and acceleration additionality while accepting that there might be more permanent effects within the umbrella of behavioural additionality.

According to this definition behavioural additionality is not a persistent effect; it operates on only one point in time during the project. Nothing spills over and endures beyond the duration of support and its immediate vicinity.

It is clearly seen from the papers in this definition that behavioural additionality is perceived as confined to R&D and innovation activities of the term as well as the temporal limitation.

B. Behavioural Additionality as the change in the non-persistent behaviour related to R&D and innovation activities:

The second group of articles that define behavioural additionality sees the concept as the change in behaviour of the agents. This change, contrary to the Category A, is beyond an extension of input and output additionality. It not only includes scale, scope and accelaration additionalities, but the way the project undertaken is also a subject of behavioural additionality.

The original definition of the concept is the prima example, Buisseret et al. (1995:590) coined behavioural addionality as “the change in a company’s way of undertaking R&D which can be attributed to policy actions.” Later, Georghiou (2002b:59) eloborates their definition by arguing that what they were inspired by while coining behavioural additionality was not the change in the “stop-go decision by the firm in respect of the project but […] rather the way in which the project was carried out.” Similarly, Georghiou (2004:7) defines it as “the difference in firm behaviour resulting from the intervention”. Clarysse et al. (2006) and Steurs et al. (2006:6) endorse this and use it to reinforce the use of the black-box analogy – behavioural additionality is what is inside the black-box left alone in between input and output additionality. Finally, Hall and Maffioli (2008: 173) use this definition in their empirical investigation.

Similar to the definition category A, the studies in this category B do not imply any persistency. The change in the behaviour does not need to endure beyond the project or its immediate vicinity. Clarysse et al. (2006) and Steurs et al. (2006:6) hint at the persistence but, as it shall be discussed later on, their temporal understanding is still more short-term than the definition categories C and D below. In a similar vein, Georghiou (1998: 39) and Davenport et al. (1998: 56) accept that behavioural additionality is "the most durable" amongst the three types of additionalities but not quite enough compared to the next two categories. Secondly, this category is also confined to the behaviours related to R&D and innovation activities.

C. Behavioural Additionality as the change in the persistent behaviour related to R&D and innovation activities:

The third category of definitions of the concept of behavioural additionality is very similar to the second one with the only difference being the element of persistence.

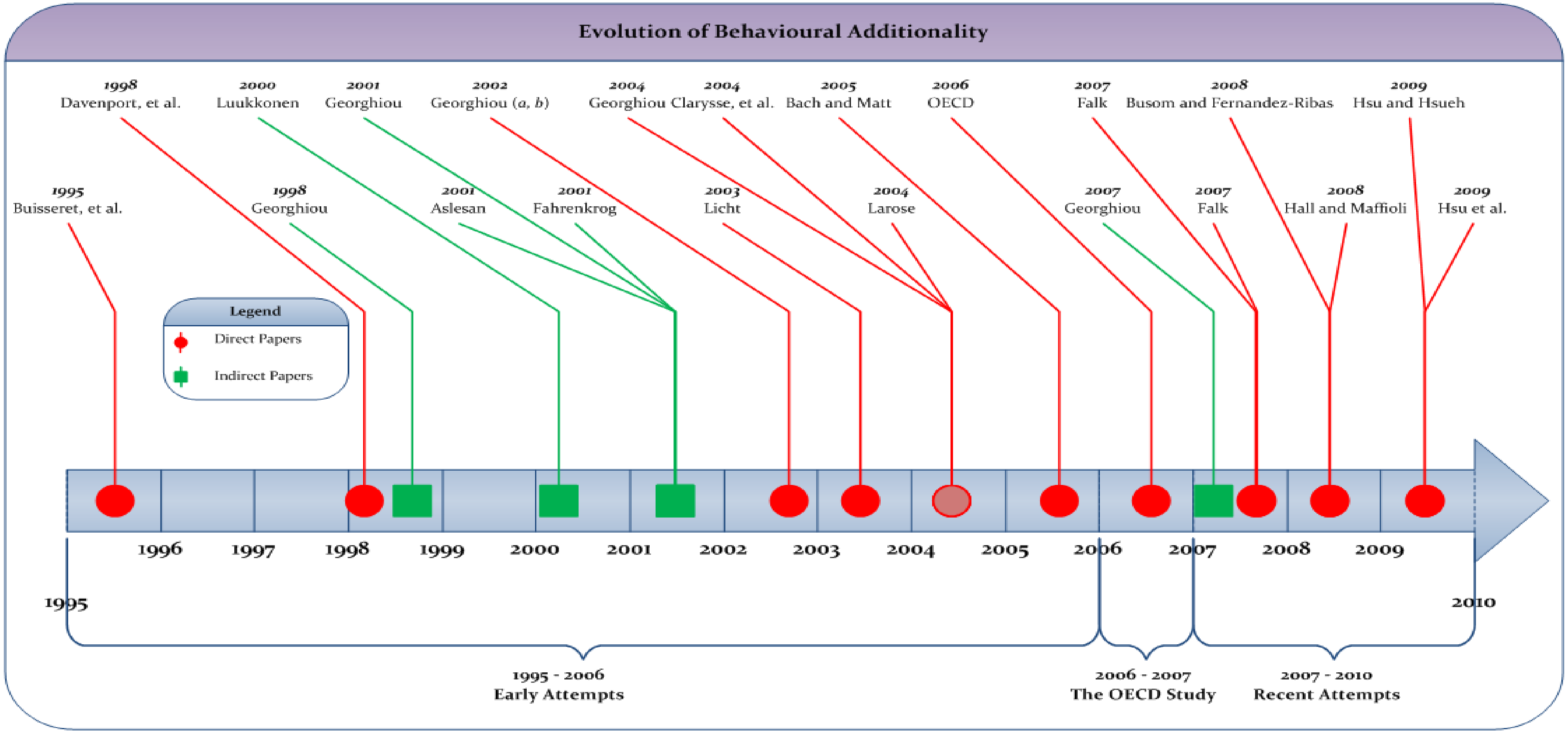
Aslesen et al. (2001:5-6) define it as the "permanent change", Licht (2003) as the change "permanent in character", and OECD (2006:187-189) as the "more sustained effects". Fier et al. (Fier et al., 2006: 127) prefer to use "long-term behaviour". Busom and Fernandez-Ribas (2008) define it as the change in the propensity to exhibit a particular behaviour. In all these definitions, persistence is the key point; these effects endure beyond the support period.

D. Behavioural Additionality as the change in of the general conduct of the firm:

The first feature in this category is that the change is not necessarily confined to R&D and innovation related activities but behavioural additionality is defined as the change in the general conduct of the firm. Secondly, behavioural additionality is defined in its widest temporal breadth and endures long after the support.

Most importantly, the definitions of behavioural additionality include more structural changes as they refer to the change in the building blocks of behaviour. The most explicit attempt to do this is by Georghiou and Clarysse (2006:12-13) who employed a resource based view of the firm and implied that behavioural additionality refers to changes in dynamic capabilities. Nonetheless, the effort was not enough to present a coherent and extensive framework as discussed earlier in this Chapter. Another attempt is by Bach and Matt (2005:37) who defined a new category of cognitive capacity additionality. Although they put this type of additionality as a fourth kind by defining behavioural additionality in category B, it is clear that what they refer is considered as part of behavioural additionality. Indeed later, Hyvarinen and Rautiainen (2007: 206) adopted this approach while defining behavioural additionality as "how public R&D funding affects the firm's behavior, cognitive capacity and learning". Finally, some other scholars defined behavioural additionality as the change in organisational routines (Georghiou, 2007; Georghiou and Keenan, 2006).

Exhibit 42: The Evolution of the Concept of Behavioural Additionality[78]


Evolution of Behavioural Additionality
1998 Davenport, et al.
2000 Luukkonen
2001 Georghiou
2002 Georghiou (a, b)
2004 Georghiou
2004 Clarysse, et al.
2005 Bach and Matt
2006 OECD
2007 Falk
2008 Busom and Fernandez-Ribas
2009 Hsu and Hsueh
1995 Buisseret, et al.
1998 Georghiou
2001 Aslesan
2001 Fahrenkrog
2003 Licht
2004 Larose
2007 Georghiou
2007 Falk
2008 Hall and Maffioli
2009 Hsu et al.
Legend
Direct Papers
Indirect Papers
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
1995 - 2006
Early Attempts
2006 - 2007
The OECD Study
2007 - 2010
Recent Attempts

---

[78] Taken from (Gok, 2010).

Table 79: Classification of Behavioural Additionality Definitions in the Academic Literature[79]

	Category A	**Category B**	**Category C**	**Category D**
Definition	an extension of input additionality covering scale, scope and acceleration additionalities and like	the change in the non-persistent behaviour related to R&D and innovation activities	the change in the persistent behaviour related to R&D and innovation activities	the change in of the general conduct of the firm substantial reference to the building blocks of behaviour
Coverage	Only R&D and innovation	Only R&D and innovation	Only R&D and innovation	Beyond R&D and innovation
Persistence	One-off, no persistence	One-off, no persistence OR Rather mid-term than long-term and rather less persistent	Persistent OR Rather long-term than short-term and rather more persistent	Persistent
Sources	(Luukkonen, 2000) (Hsu et al., 2009; Hsu and Hsueh, 2009) (Malik et al., 2006) (Georghiou, 2002b)	Buisseret et al. (1995:590) Georghiou (2002b:59) Clarysse et al. (2006) Steurs et al. (2006:6) Hall and Maffioli (2008: 173) Georghiou (1998: 39) Davenport et al. (1998: 56)	(Lenihan et al., 2007: 317-318) Aslesen et al. (2001:5-6) Licht (2003) OECD (2006a:187-189) (Fier et al., 2006: 127) (Busom and Fernandez-Ribas, 2008: 241)	(Georghiou, 2007: 744) Bach and Matt (2005:37) Georghiou and Clarysse (2006:12-13) (Georghiou and Keenan, 2006: 770) (Hyvarinen and Rautiainen, 2007: 206)

---

[79] Source: Gok (2010)

3 Statistical Analysis: What differs in evaluations with behavioural additionality

3.1 Introduction

Among other characteristics, the INNO-Appraisal investigated whether evaluations covered particular evaluation topics, i.e. questions that the evaluation aims to answer. As shown in the overall analysis of our dataset in Chapter 3 of this report, out of 15 different topics of evaluation, three of them are input additionality (OA), output additionality (IA) and behavioural additionality (BA) – the three types of additionality introduced in the previous section. As the main aim of the analysis is to explore the distinct features of additionality, and more concretely of behavioural additionality, this part will look at how the coverage of those types of additionality is linked to other evaluation characteristics.

As established from the review of scholarly literature on this topic above and elsewhere (Gok, 2010), the difference between the traditional types of additionality (input and output additionality) and behavioural additionality is very vague as i) these three concepts are often intertwined and also ii) various approaches to behavioural additionality define various relationships between these three additionality concepts. Therefore, throughout the Chapter all three concepts of additionality will be analysed, enabling individual and comparative analysis of BA. The idea, as a first step, is to see if the evaluations covering input, output and behavioural additionality are statistically significantly different than the whole set (and hence from the evaluations that do not cover any type of additionality) in terms of the following evaluation characteristics.

- Timing of evaluation
- Purpose of evaluation
- Budget, planning and sponsorship and tendering of evaluation
- Impacts looked at in evaluation
- Main evaluation designs
- Main data collection methods and data sources employed in evaluation
- Main data analysis methods used in evaluation
- Main intended audiences for evaluation
- Terms of reference availability
- Quality of evaluation
- Usefulness of recommendations of evaluation
- Discussions of evaluation
- Consequences of evaluation

Methodologically, for categorical cross-tabulations a Chi-Square test at 90% confidence will be employed, while for correlations the Pearson or Spearman test at 90% confidence will be used. All the significant associations and correlations will be indicated.

3.2 Coverage of Additionality

As outlined in Exhibit 43, about 50% of evaluations covered behavioural additionality, output additionality and input additionality. The first observation here is that behavioural additionality has gained a place for itself. Although it is quite a new concept compared to the more established concepts of output and input additionality, the uptake of the former is not less than the latter ones. Secondly, in spite of the fact that there are clearly more popular evaluation topics than (any kind of) additionality, they are not marginal or outlier topics in terms of their uptake.

Exhibit 43: Uptake of Additionality as a Topic of Evaluation

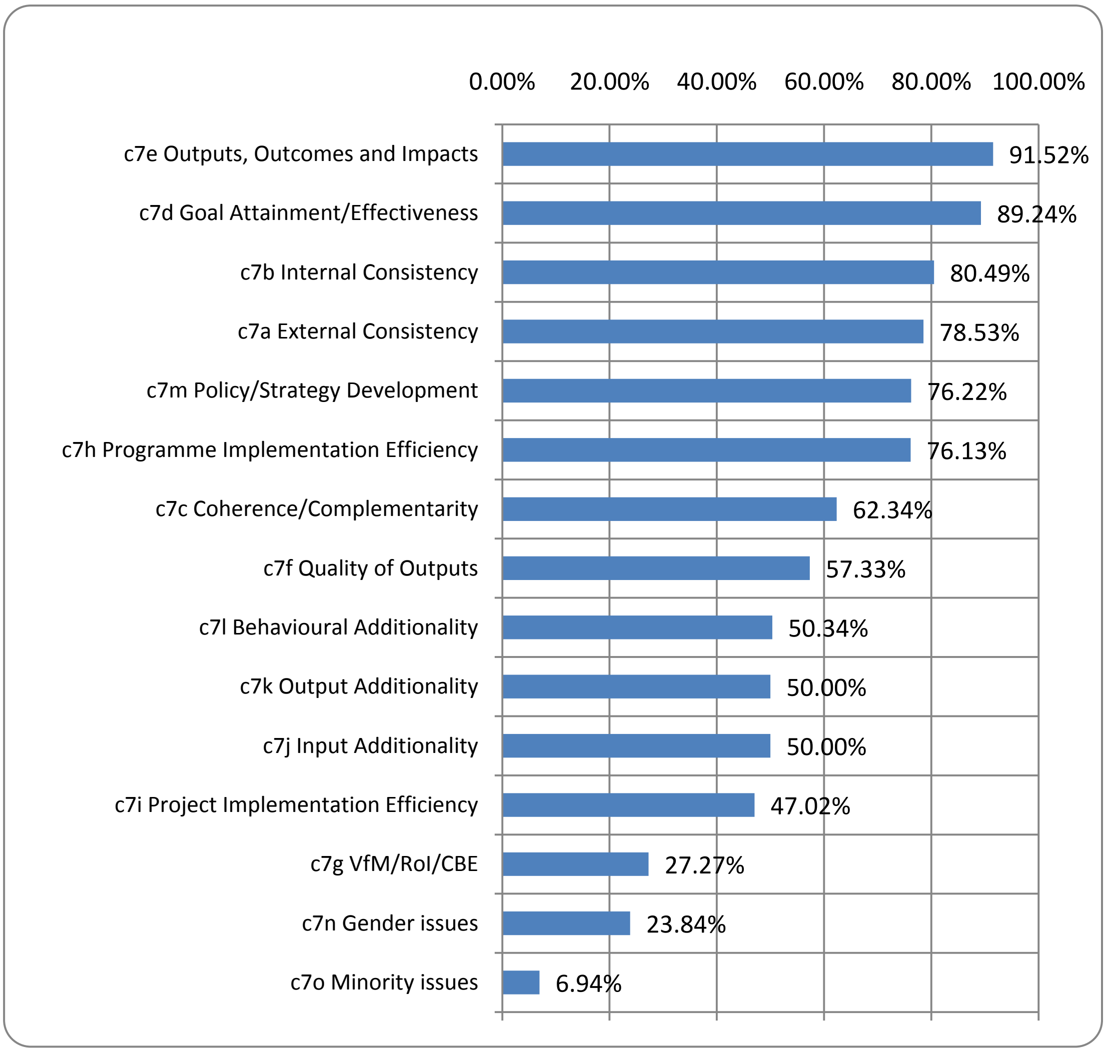


Exhibit 44 shows the relationship between the evaluations covering the three types of additionality and also the remaining evaluations that do not cover any type of additionality. Two-thirds of the national innovation policy measure evaluations in the database cover at least one form of additionality, and one-third of evaluations cover all three types of additionality. Those covering exclusively behavioural additionality constitute only one-fifth of all behavioural additionality evaluations. This picture, therefore suggest that the three types of additionality are used extensively and they are predominantly used together.

Exhibit 44: Euler Diagram of the Coverage of Types of Additionality in Combination

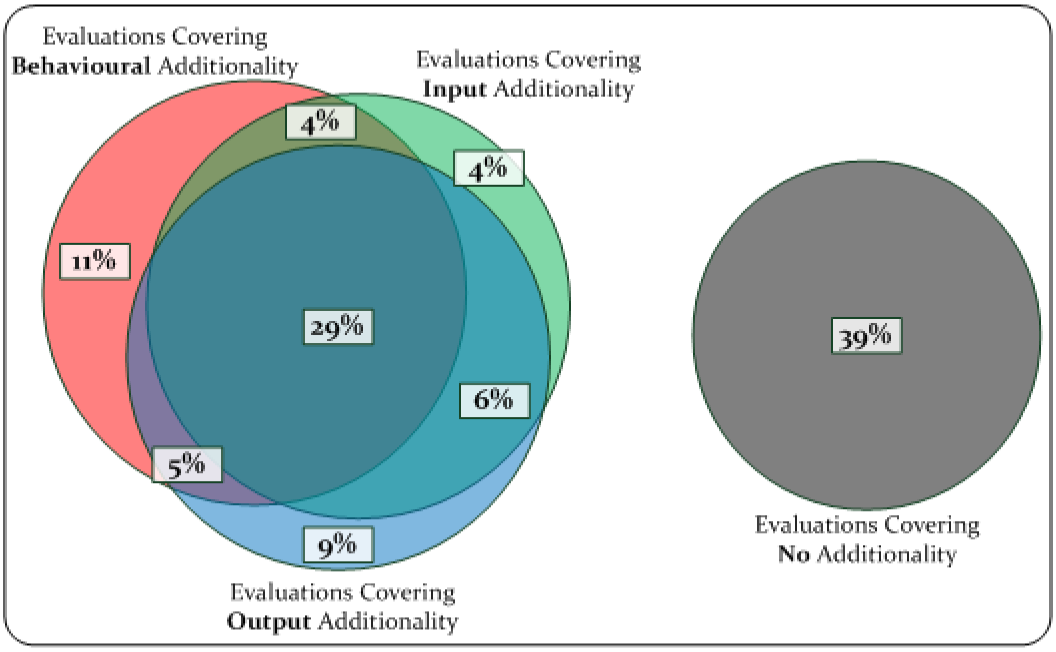
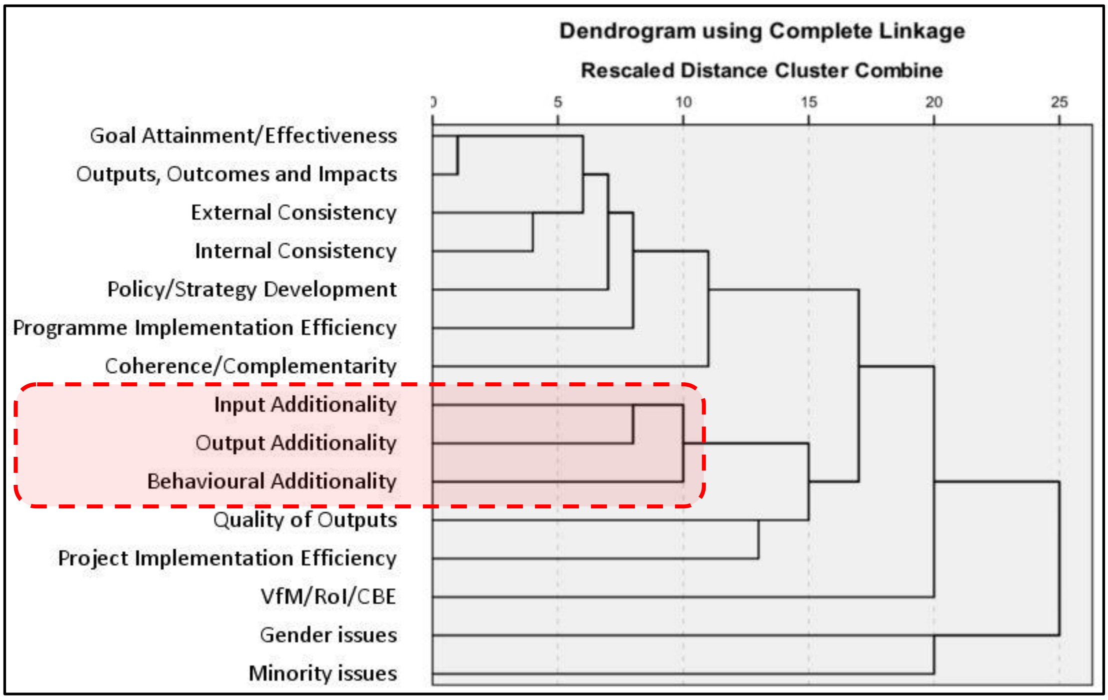

Evaluations Covering
Behavioural Additionality
Evaluations Covering
Input Additionality
4%
4%
11%
29%
6%
5%
9%
Evaluations Covering
Output Additionality
39%
Evaluations Covering
No Additionality

Exhibit 45 further confirms this picture. It shows the illustration of the results of the clustering analysis of the evaluation topics. As it can be read from the dendrogram, input, output and behavioural additionality clearly form a cluster – they tend to be used together in evaluations. These three types of additionality are mostly linked with the cluster of topics formed by “quality of outputs” and “project implementation efficiency”, while there is no close link to “programme implementation efficiency”. This is not surprising considering the fact that behavioural additionality is especially linked with the micro level, the firm or even project level, the way a project is conducted or actions and routines are changed, and with the immediate difference this makes for the output (quality). This also points to the differences between evaluations that are more closely related to the project level and those that are more interested in the programme level efficiency, as the two are not closely linked.

Interestingly, the cluster formed by the three types of additionality is not very closely related to topics such as “outputs, outcomes and impacts”. This seems to indicate that the additionality dimensions are not simply add-ons to the traditional output and impact dimension, but often used independently.

Exhibit 45: Dendrogram of Topics Covered (Average Linkage Between Groups)[80]



3.3 Measure Types and Target Groups

As depicted in Exhibit 46, only a small fraction (4.70%) of the evaluations in the INNO-Appraisal database is for tax incentives. However, for all three types of additionality, it is slightly but statistically more significantly.

The biggest category of policy measures is “direct financial support for innovation activities” with almost 60% of all evaluations; and this share is similar in the three subsets.

Innovation management measures account for around one-third of the sample and while behavioural and output additionality subsets are close to this figure, only one-fifth of the evaluations of such measures covered input additionality. Similarly, around 7% of the evaluations are associated with measures aiming development and creation of intermediary bodies and this ratio is statistically significantly lower for input additionality while it is close to average for behavioural and output additionality.

Mobility of personnel measures are about 8% of the whole set and also the three subsets. Similarly, measures targeting the creation of start-ups are 8% but in this case these measure’s evaluations cover more input additionality (around one-seventh).

---

[80] Rezankova (2009) recommends “Jaccard’s co-efficient” or “Yule’s Q” measures for object clustering (clustering of variables of same type) of dichotomous (variables that take binary options) asymmetric (“1” and “0” values are of inherently different importance) variables. This method does not cluster variables on the basis of co-absence of same trait (i.e. both variables takes the value “0” at the same time). In this analysis, furthest neighbour method which links topics with complete linkage is used by applying Jaccard’s co-efficient measure.

Measures aiming at the creation of networks, clusters and collaboration are around one third of the sample and input additionality subset. They account for around 40% of output and behavioural additionality subset and the difference is statistically significant. Science and industry cooperation measures, however, are not different for the three subsets than the whole set – they account for around a quarter of the data.

Finally, support for uptake and diffusion type of measure are around one-fifth and only the input additionality subset differs here by being around one-seventh.

To sum up, behavioural additionality evaluations are predominantly for the measures aiming direct financial support (around 54%), innovation management support and dissemination, innovation culture (around 33%) and Networks & Clusters, collaboration and Technology/Knowledge Transfer (around 40%). Furthermore, there are not many significant differences between the whole set and the behavioural additionality subset in terms of the measure they are associated with.

Exhibit 46: Measure Types and Additionality

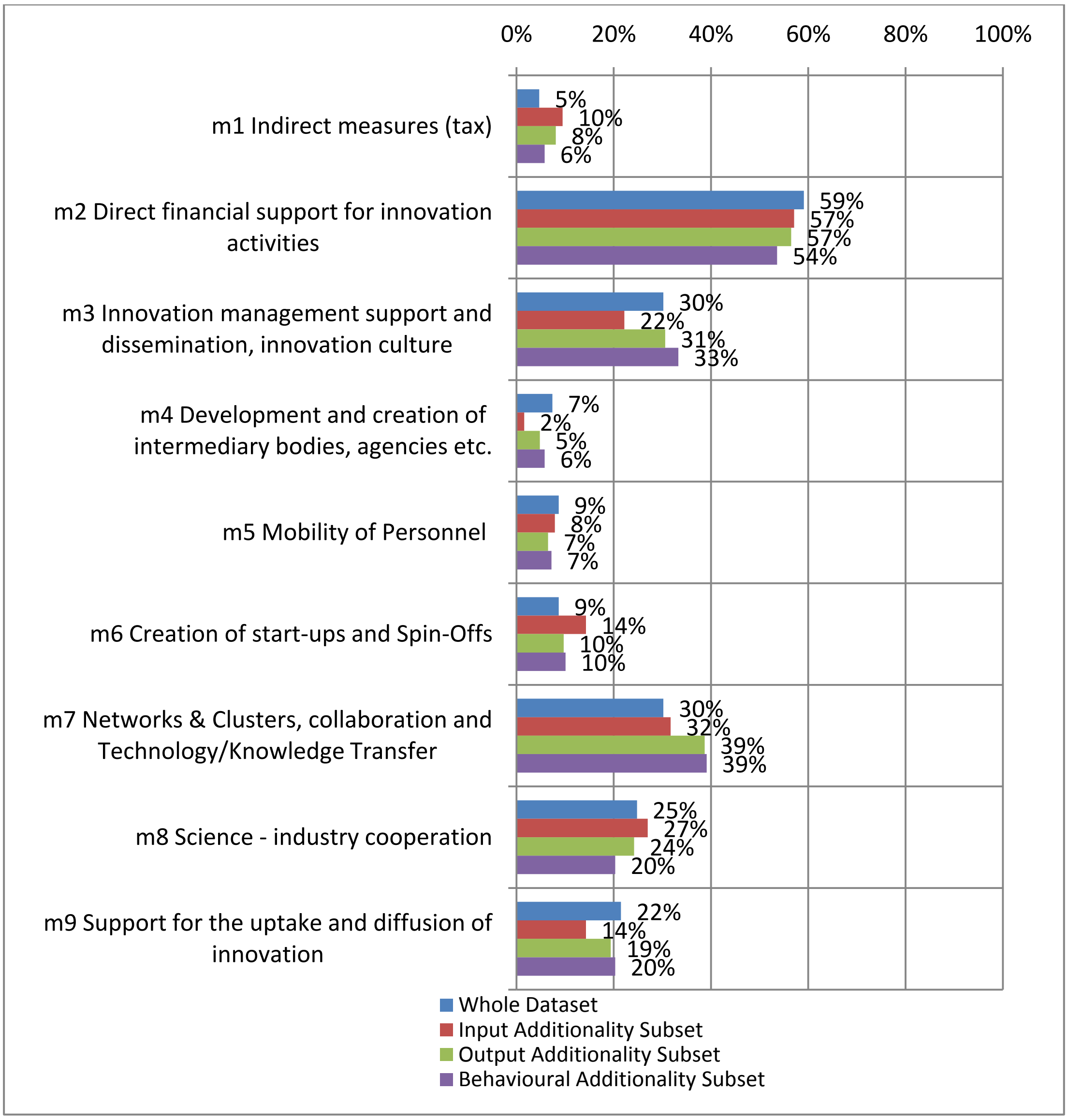
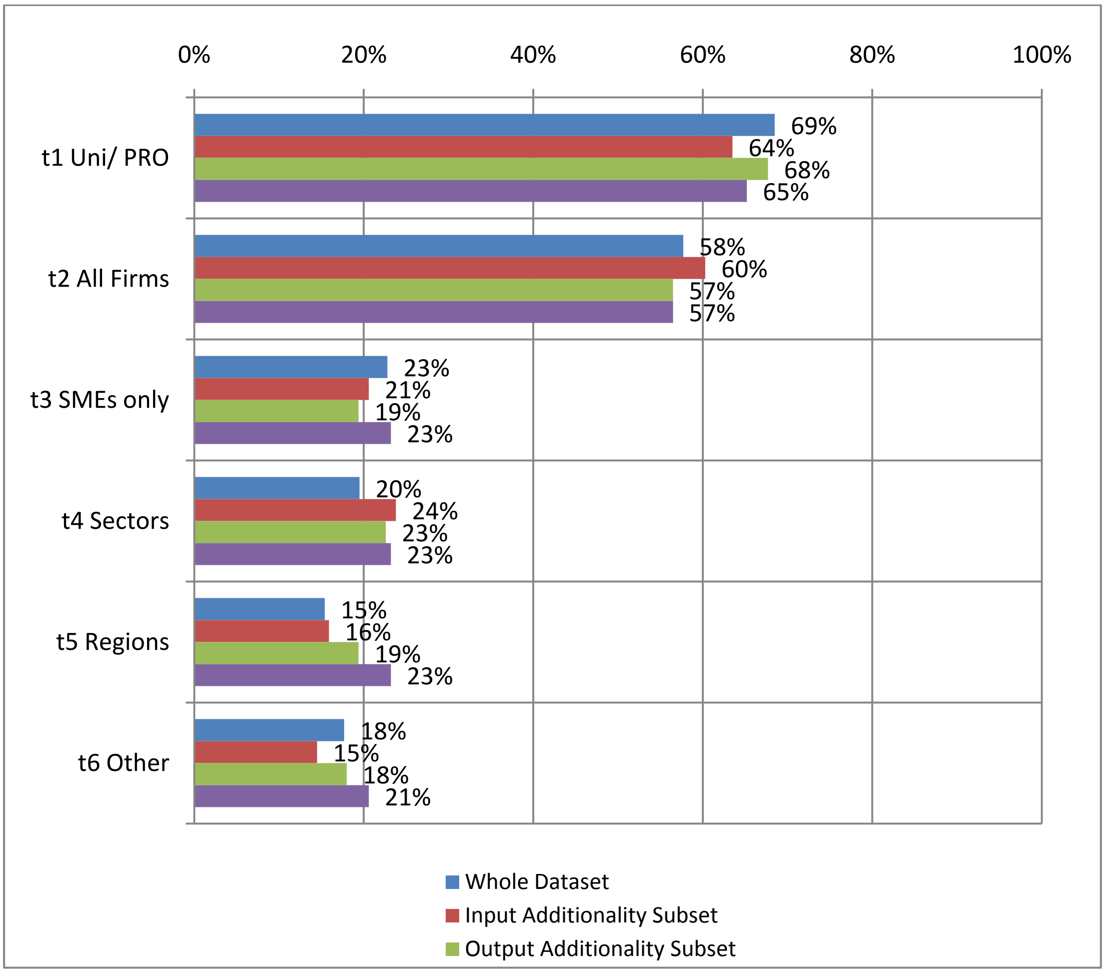


Exhibit 47 depicts that the main target groups of the measures whose evaluations are included in the INNO-Appraisal database are universities and public research organisations (around two-third), firms (58%), only SMEs (23%), sectors (20%) and regions (15%). The ratios for the three subsets are not statistically significantly different than these with the only exception that being input additionality is slightly more associated with measure targeting sectors (25%).

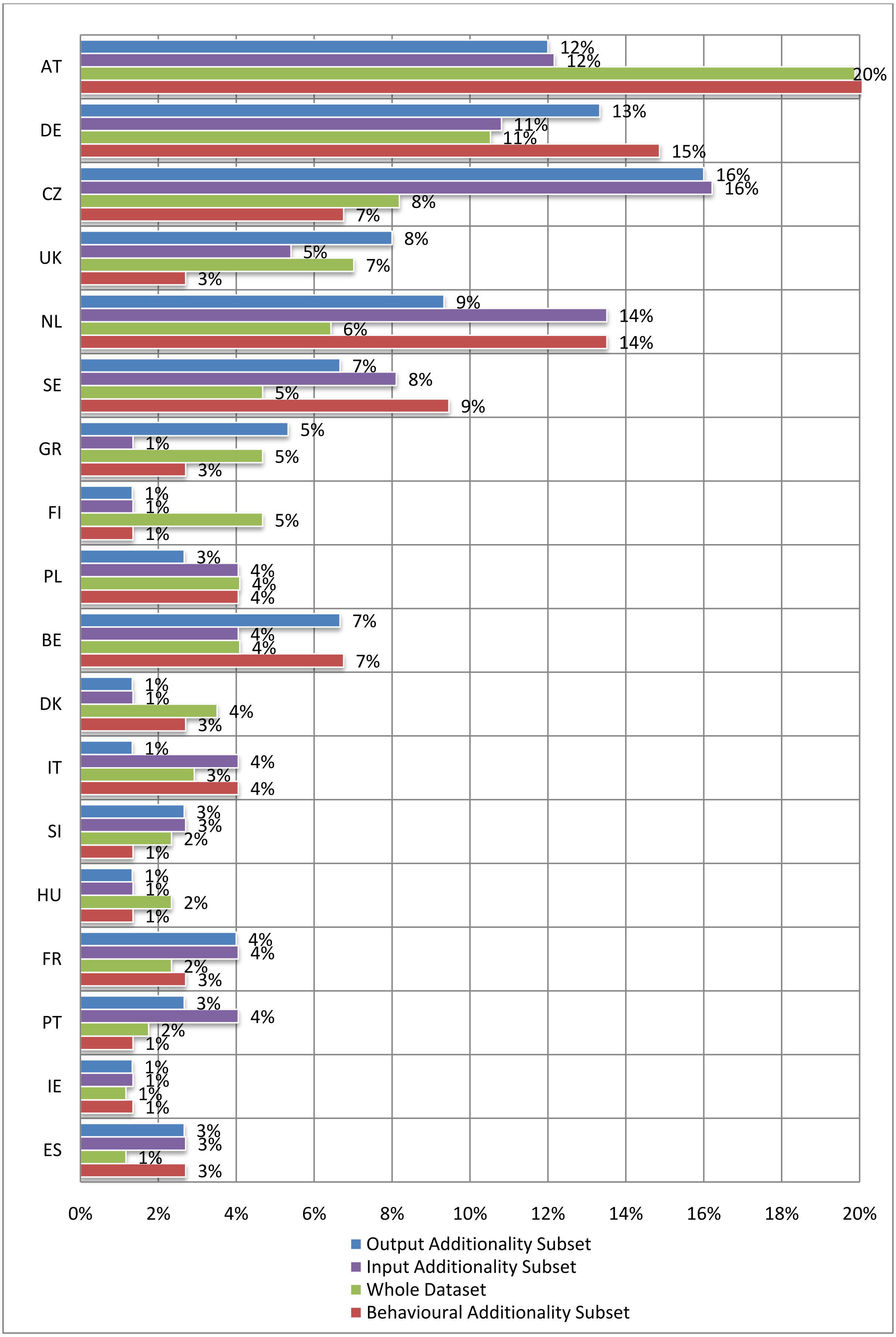
Exhibit 47: Measure Target Groups and Additionality



3.4 Country Distribution

A further analytical question is that if there are any country biases, any differences in the use of additionality issues in evaluations. As depicted in Exhibit 48 below, the distribution of the evaluations according to countries shows that input, output and behavioural additionality were covered more in the evaluations of Austria, Germany, Czech Republic, the UK, and the Netherlands. As these countries are, at the same time, among those that show the highest evaluation activity level, there appears to be a link between having high evaluation activity level on the one hand and using behavioural additionality as a concept on the other hand.

Exhibit 48: Country Distribution of Evaluations and Additionality



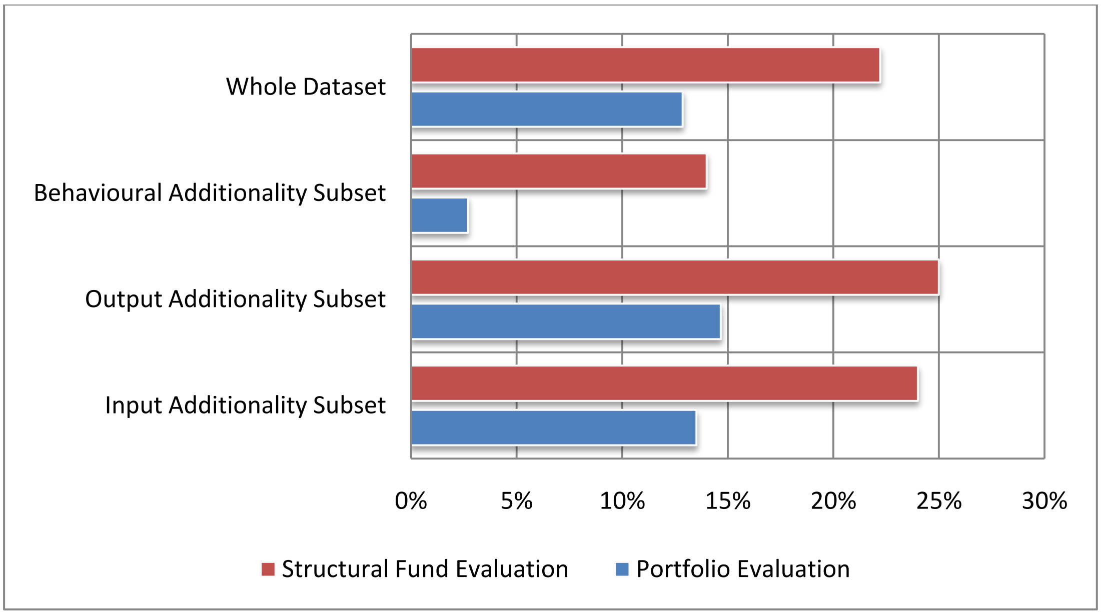
3.5 Structural Funds and Portfolio Evaluations

As depicted in Exhibit 49, evaluations of measures supported by the EU Structural funds, i.e. structural fund evaluations, form a significant subset - around 20% - in the INNO-Appraisal dataset. These evaluations have significantly different characteristics compared with other evaluations (see Chapter 5.3 for an in-depth analysis of structural fund evaluations).

Similar to structural fund evaluations, the set of evaluations that covered more than one policy measure in one report, portfolio evaluations, is a significant and distinct subset with circa 13% of the whole dataset. Most of the portfolio evaluations are either structural fund evaluations or they belong to Austrian measures.

This distribution is more or less the same for input and output additionality evaluations while for behavioural additionality the share of structural fund and portfolio evaluations is statistically significantly lower than the overall, around 15% and 3% respectively. This might suggest that as structural fund evaluations are imposed by the European Union and they investigate the strategic impact in the macro level, behavioural additionality – a topic that is used as policy and operational learning and a topic that is generally considered as micro – is not embraced by structural fund evaluations.

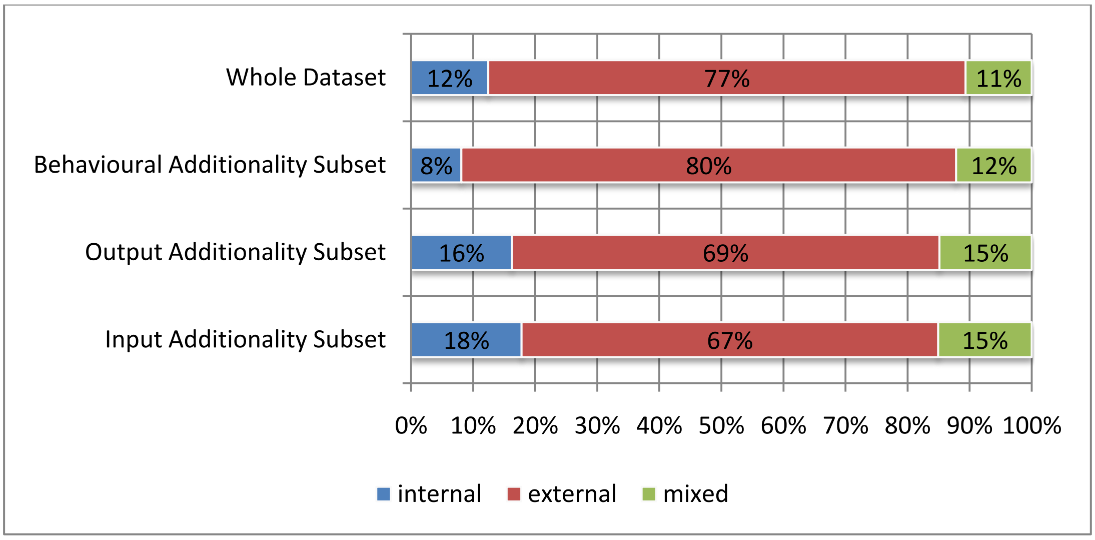
Exhibit 49: Structural Fund and Portfolio Evaluations and Additionality



3.6 Evaluator and Tender Process

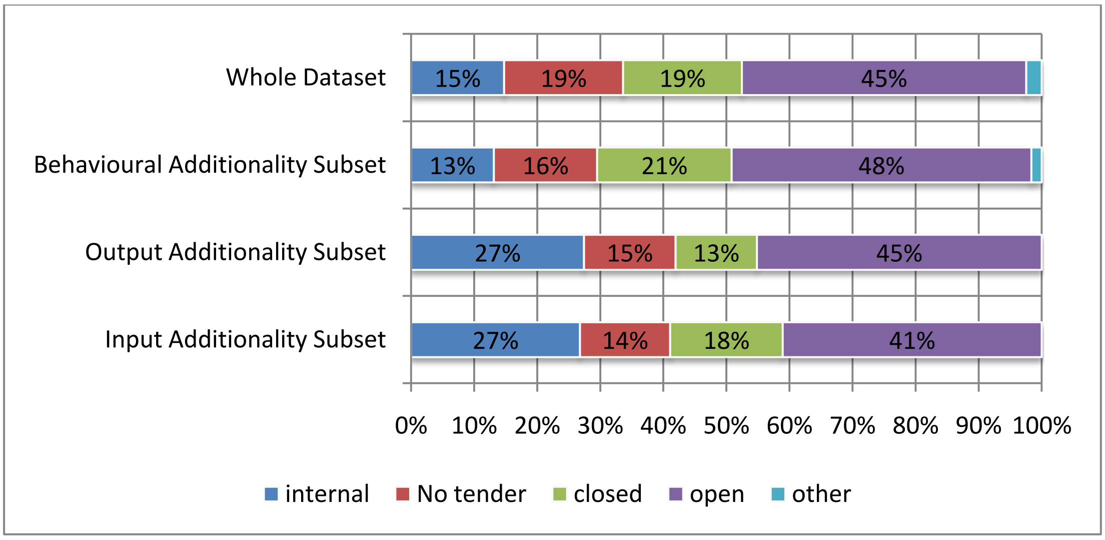
As shown in Exhibit 50, around three quarters of all evaluations in the sample were conducted by external evaluators while circa 15% were conducted by internals and around 11% were conducted by both internal and external evaluators. The distribution is similar for all three types of additionality and there are no statistically significant differences. This suggests that behavioural additionality is not understood and embraced by internal evaluators as much as external evaluators.

Exhibit 50: Type of Evaluator and Additionality



The tender processes of the evaluations in the INNO-Appraisal sample are dominantly open (circa 45%) as depicted in Exhibit 51. Internal, no tender and closed tender processes share the other half of the sample equally (around 17% each). For those evaluations that covered behavioural additionality, the distribution is more or less the same (open tender is slightly higher but it is not statistically significant). For input and output additionality, however, internal tender process has a statistically significantly higher share at the expense of no tender and closed processes. This might be due to the fact that input and output additionality evaluations are generally perceived as hard-to-conduct as to their methodologies, which involve sophisticated econometric techniques with quasi experimental designs. In non-closed tender processes (internal and open), it might be perceived that the choice of evaluator is limited to those quantitatively focused evaluators that lack the context knowledge and thus the ones that have the context knowledge but lack the sophisticated and sometimes experimental econometric knowledge are excluded. Therefore, the issuers of evaluation contracts might need to control the process against this kind of bias by making the tender process closed. Another explanation can be that internal evaluators are not as familiar with the BA concept as with other concepts, which also confirms the result for the choice of evaluator.

Exhibit 51: Tender Process of Evaluation and Additionality



3.7 Timing and Purpose

The timing of the evaluations in the INNO-Appraisal sample is depicted in Exhibit 52. Almost half of the evaluations (43%) are accompanying, which is defined as evaluations conducted during the course of the policy measure but multiple points in time. Ex-ante evaluations are circa 13% of all cases. Similarly interim evaluations which are different from accompanying evaluations by being conducted one point in time are also 14%. Finally, ex-post evaluations are circa 28%. The dataset is slightly biased against the ex-post evaluation owing to it initially relying on the INNO-Policy Trendchart database to identify the evaluated policy measures, which only includes running policy measures / programmes.

The distribution of timing options is statistically significantly different for all three types of additionality. Those evaluations which cover any type of additionality are more often ex-post, as often accompanying and less often ex-ante and interim. This points to the medium and long term nature of additionality effects. Further, and more interestingly, behavioural additionality is not linked with accompanying evaluations as much as one would expect. A good accompanying evaluation should rely on behavioural additionality based on monitoring data, as this would allow to re-engineer and re-enforce desired effects while the programme is running. Therefore, it can be argued that real-time evaluation dimension of behavioural additionality is still under-explored.

Exhibit 52: Timing of Evaluation and Additionality

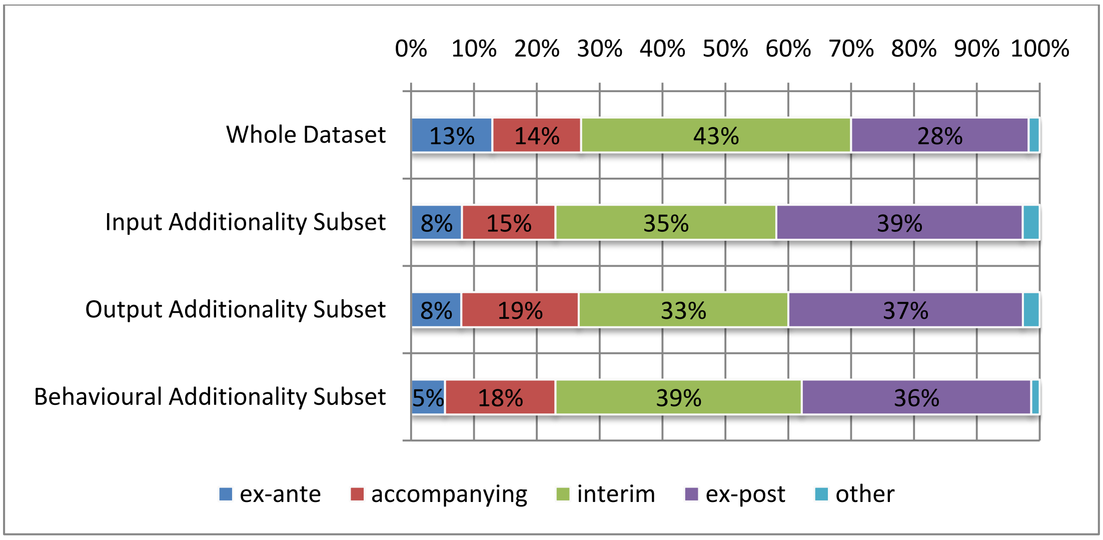
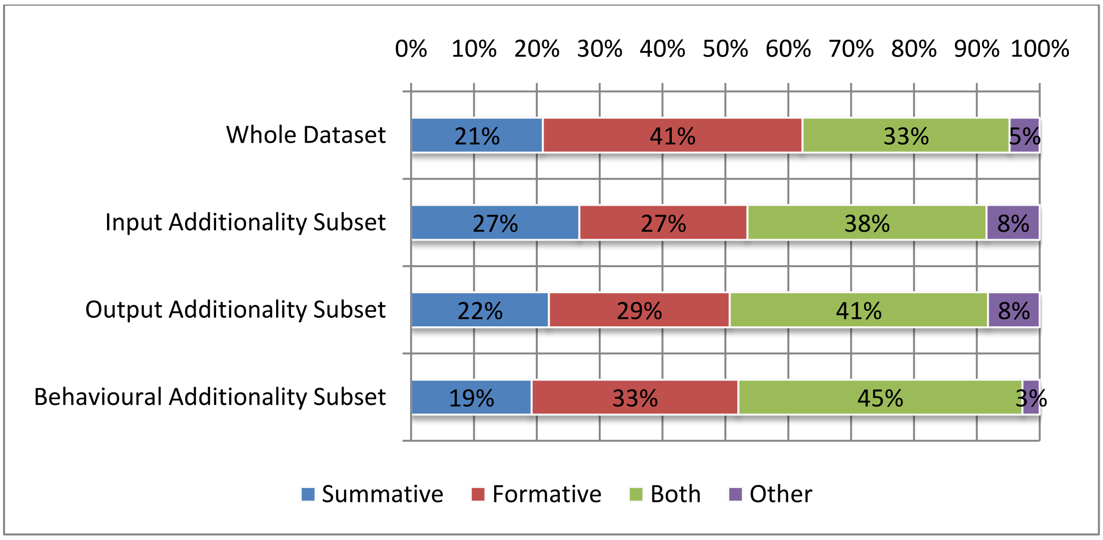


Exhibit 53 shows the purpose of evaluations in the INNO-Appraisal sample: 41% of the evaluations are formative, while summative and mixed type evaluations consist circa 21% and 33% of the sample respectively. All three types of additionality evaluations are statistically significantly different from the whole set, as they are less formative and more summative and mixed type. As per behavioural additionality, although it is less formative than the whole set, it is more formative than input and output additionality. But again, the same applies as above, behavioural additionality is not as formative as one would expect or as the concept itself offers.

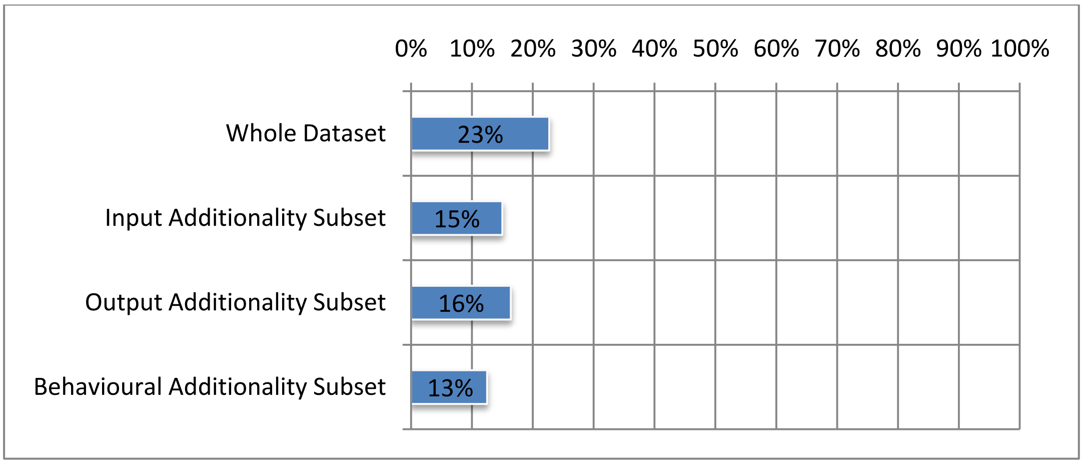
Exhibit 53: Purpose of Evaluation and Additionality



3.8 Budget and Planning of Evaluations

Exhibit 54 depicts the situation for the evaluations which were the condition of external sponsorship. In general these evaluations make up around one quarter of evaluations while for all three types of additionality this figure is statistically significantly lower, around 15%. This seems logical as more often structural fund evaluations are conditions for EU funding and the share of structural funds evaluations is lower in additionality covering evaluations.

Exhibit 54: Condition of External Sponsorship and Additionality



As shown in Exhibit 55, nearly half of the evaluations had a dedicated budget from the very beginning of the measure's design. This ratio is similar in all three types of additionality.

Exhibit 55: Dedicated Evaluation Budget and Additionality

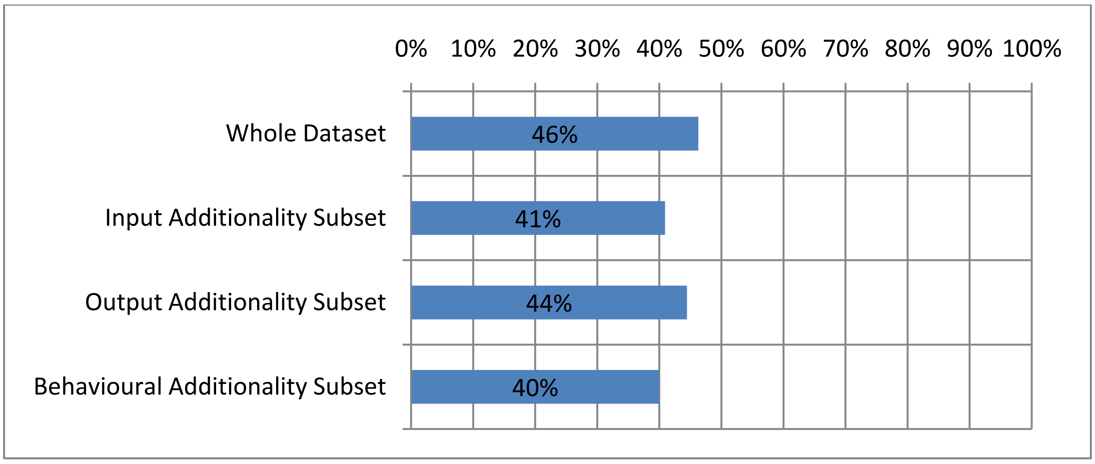
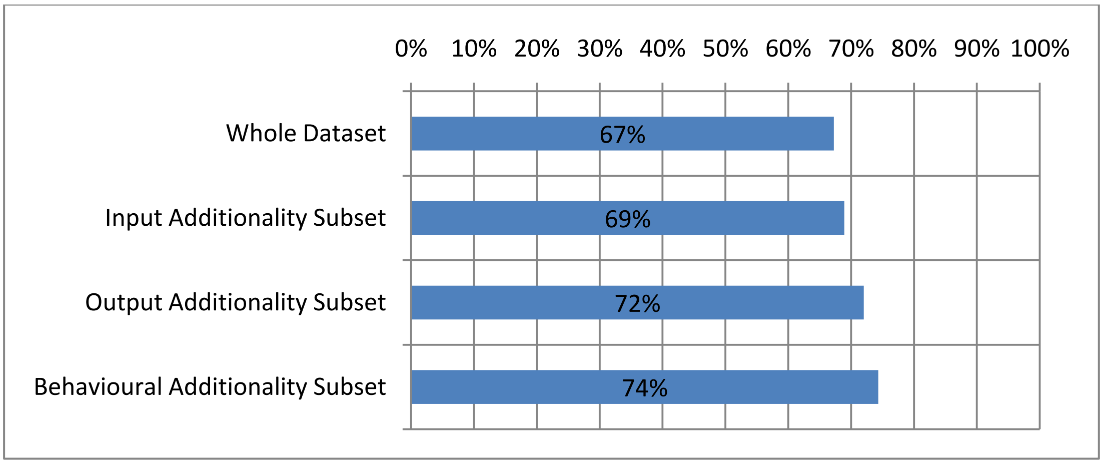


Exhibit 56 shows the share of foreseen and planned evaluations . Around two-thirds of evaluations were foreseen and planned from the very beginning and this ratio is more or less similar in input, output and behavioural additionality.

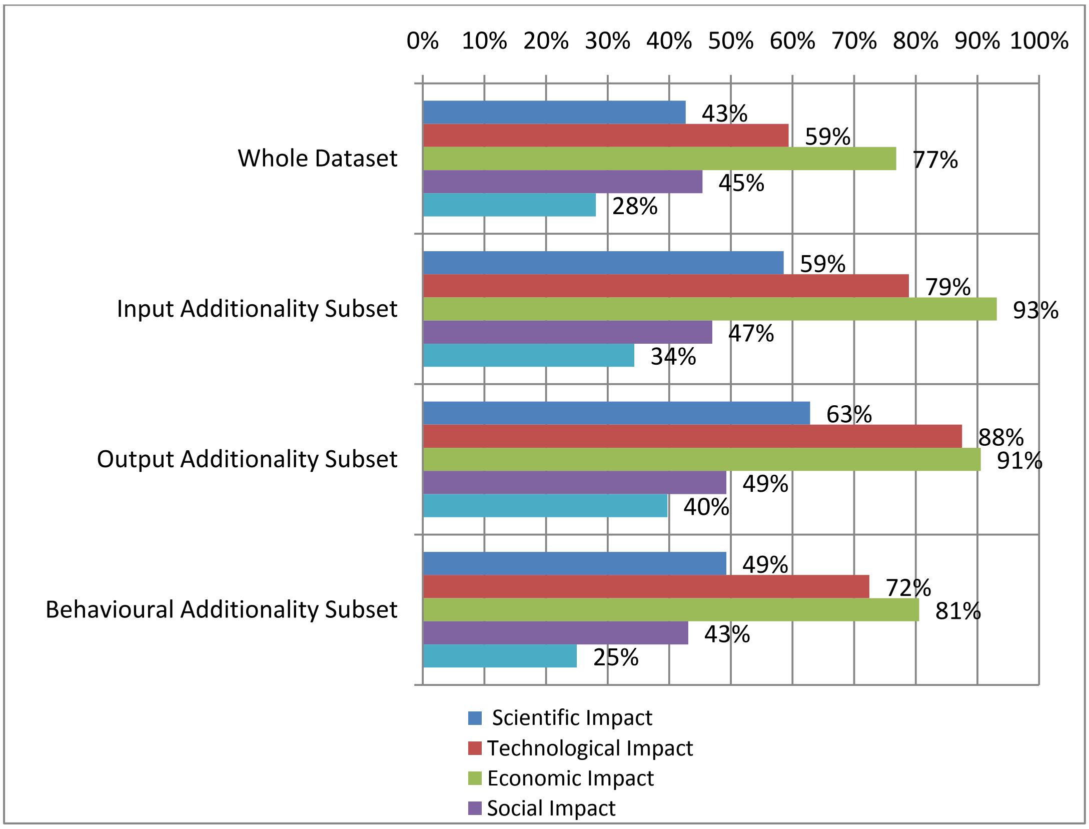
Exhibit 56: Foreseen and Planned Evaluation and Additionality



3.9 Impacts Looked at in Evaluations

The analysis of the type of impact looked at in evaluations yielded interesting results. For the whole sample, circa 43% of evaluations looked at scientific impacts (Exhibit 57). This ratio is statistically significantly different for all three types of additionality: those evaluations that covered additionality looked at scientific impact more than the whole sample. For behavioural additionality it is less significant than the other two as the difference between the share of evaluations that looked at scientific impacts in the whole sample and behavioural additionality is only 6% while for the other two types of additionality, it is around 20%.

Exhibit 57: Impacts Looked at in Evaluations and Additionality



Technological impacts were looked at 59% of the whole sample. The situation is more statistically significantly different for additionality subset from the previous case: evaluations covering all three types of additionality looked at technological impacts more often. The ratio of the evaluations that looked at technological impacts within behavioural additionality, input additionality and output additionality subsets are 72%, 79% and 88% respectively.

The situation with economic impacts is interesting. For the whole set, the share of evaluations looking at economic impacts (77%) are more than those looking at scientific (43%) and technological impact (59%). Furthermore, this share is even larger for input (93%) and output (91%) additionality subsets, and these differences are statistically significant. However, for behavioural additionality the ratio that looked at economic impacts (81%) is not statistically different than the whole set. This suggests that behavioural additionality is not currently linked as strongly with economic value creation as input and output additionality.

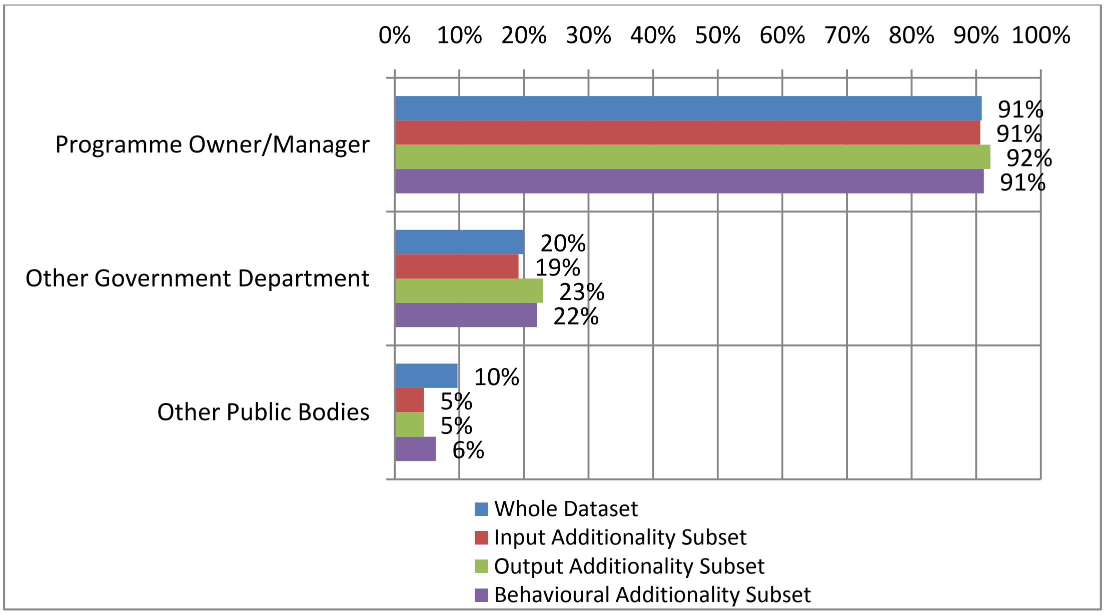
Around 45% of the whole sample looked at social impact, and all three types of additionality had similar percentages. Finally, for environmental impacts, the ratio is only 28% for the whole set which makes environmental impacts the least popular impact type. The results for the behavioural additionality subset are not statistically different from the larger set. However, output (40%) and input (34%) additionality have higher ratios and these differences are statistically significant. Again, this yields interesting results: behavioural additionality is not particularly linked with social impacts and with environmental impact, while at least the latter is strongly associated with input and output additionality.

To sum up, around half of behavioural additionality evaluations looked at scientific impact, around three quarters at technological impacts, around 80% at economic impacts, around half at social impacts and only a quarter at environmental impacts. For scientific and technological impacts these ratios are statistically significantly higher than the whole dataset.

3.10 Sponsors of Evaluation

Exhibit 58 shows that more than 9 evaluations in every 10 have the programme owner / manager as their sponsor while only one-fifth of the whole set are sponsored by another governmental department. For the three subsets these numbers are more or less the same, indicating that there are no significant biases against, or in favour of, additionality aspects in any of the potential sponsors; additionality is not imposed by external sponsors.

Exhibit 58: Sponsors of Evaluation



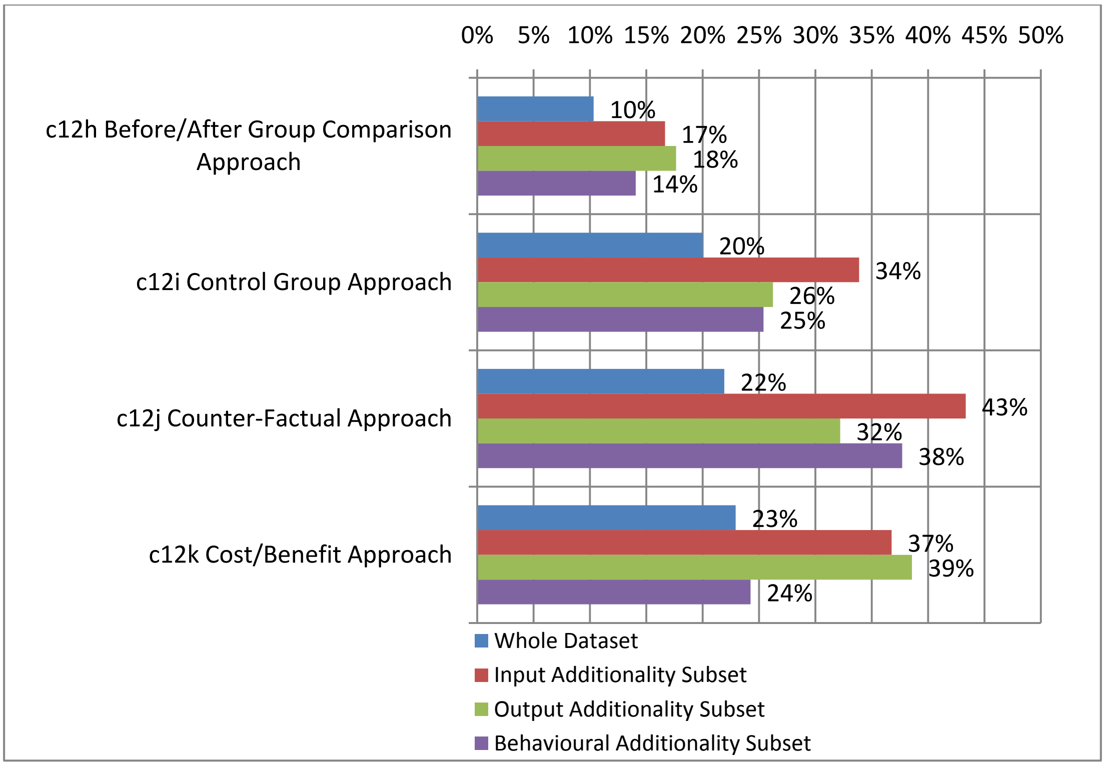
3.11 Evaluation Designs

Exhibit 59 depicts the (quasi-experimental) evaluation designs: within the whole sample only one quarter employed a cost-benefit approach, approximately one-fifth used counter-factual and/or control group approach and only one-tenth utilised before-after group comparison.

The figures are statistically significantly higher for input and output additionality which employed these approaches heavily more than the whole set - around three quarters of these two subsets employed these designs. The margins are higher for input additionality than output additionality.

For behavioural additionality the situation is somewhat puzzling. Despite the percentage of evaluations employing these designs in behavioural additionality is slightly higher than that of the whole set, the only difference in counter-factual approach is statistically significant. This might be due to the fact that behavioural additionality evaluations rely on non-experimental designs rather than quasi-experimental ones. Similarly, while input and output additionality are closely linked with cost-benefit dichotomy, behavioural additionality is not particularly linked to it. This is another indication that behavioural additionality is not as strongly linked to immediate economic effects as the other two forms, reflecting the indirect – and underestimated – effect of change in behaviour on outcome.

Exhibit 59: Evaluation Designs



3.12 Data Collection Methods

Exhibit 60 depicts the various data collection methods. Around four-fifths of the set utilised monitoring data and this ratio is similar for the three subsets. Roughly two-thirds of all four sets employed a document search, only a tiny fraction utilised technometrics methods (as the dataset consists of innovation policy measure evaluations rather than science policy) and only one-seventh collected data through peer-reviews. Roughly half of the evaluations used a form of focus groups / workshops / meetings as the data collection method. More than three-quarters used interviews to this end. Interviews were used statistically significantly more (89%) by the behavioural additionality subset.

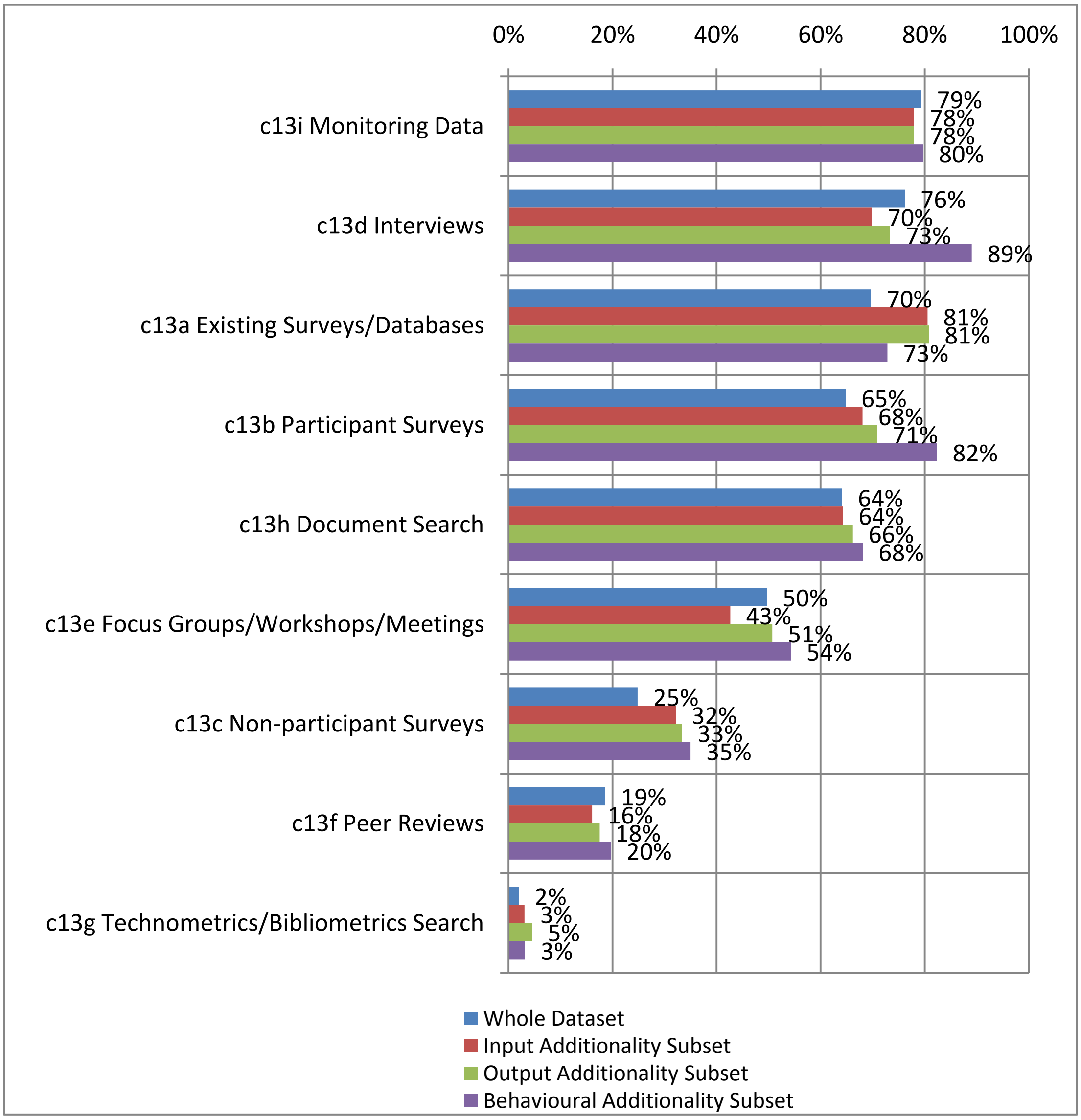
Non-participant surveys consist of a quarter of the whole data while it is statistically significantly more for input, output and behavioural additionality (around one-third). Participant surveys were employed by 65% in general, while for behavioural additionality it is around 82% and the difference is statistically significant. Finally, around 70% of all evaluations utilised existing surveys and databases; for input and output additionality this ratio was 10% higher.

To sum up, behavioural additionality utilised interviews (89%, significantly more than the others), monitoring data (80%), participant surveys (about 82%, significantly more than the others), existing surveys/databases (73%), document search (64%), non-participant surveys (35%, significantly more than the others), peer-reviews (20%) and technometrics/bibliometrics (2%) as data collection methods.

It appears that for key methods there is a striking difference not only between behavioural additionality and the whole sample, but also between behavioural additionality and the other two additionality concepts, while those two appear to be highly similar in terms of the data methods used. Most significantly in a statistical sense, behavioural additionality evaluations use more often interviews and surveys, while they relying slightly less (if not significantly) than other types of additionality, and the sample as a whole on existing survey and data.

This picture both endorses and contradicts the rather strong claim that behavioural additionality cannot be understood only and exclusively by survey based evaluations (c.f. Georghiou (2007)). On one hand, behavioural additionality evaluations need more in depth data collection practices like interviews than other evaluations. On the other hand, the majority of behavioural additionality evaluations generally utilised quantitative methods such as monitoring data, existing and new surveys. This might be due to the fact that, as discussed above, behavioural additionality is rarely evaluated exclusively and one needs complementary data collection methods to evaluate all three types of additionality separately and as a whole. Finally, the less pronounced use of existing data is logical, as existing data is rarely collected for the purpose of behavioural additionality and thus lacks important variables and dimensions, while data needed for input and output additionality is more in line with widely available statistics.

Exhibit 60: Data Sources and Collection Methods and Additionality



3.13 Data Analysis Methods

As shown in Exhibit 61, almost two-third of the INNO-Appraisal subset employed context analysis. Interestingly, although not statistically significant, this ratio is slightly lower for behavioural additionality subset. Furthermore, for input and output additionality this ratio increases up to three quarters and the difference is statistically significant. This picture clearly suggest that the concept of behavioural additionality is not employed properly, as any definition of behavioural additionality that was reviewed in above requires an explicit understanding of the context conditions.

Input and output additionality evaluations are – as to be expected – significantly more likely to use input/output analysis than the whole sample and also than behavioural additionality evaluations. Again, this confirms that behavioural additionality is less linked to concrete, measurable output.

A similar distinction between behavioural additionality on the one hand and input and output additionality on the other hand can be observed with descriptive statistics, which are used in behavioural additionality evaluations much more often than in other types of evaluations. Practically all behavioural additionality evaluations employ descriptive statistics, the ratio is 97% and the difference is statistically significant.

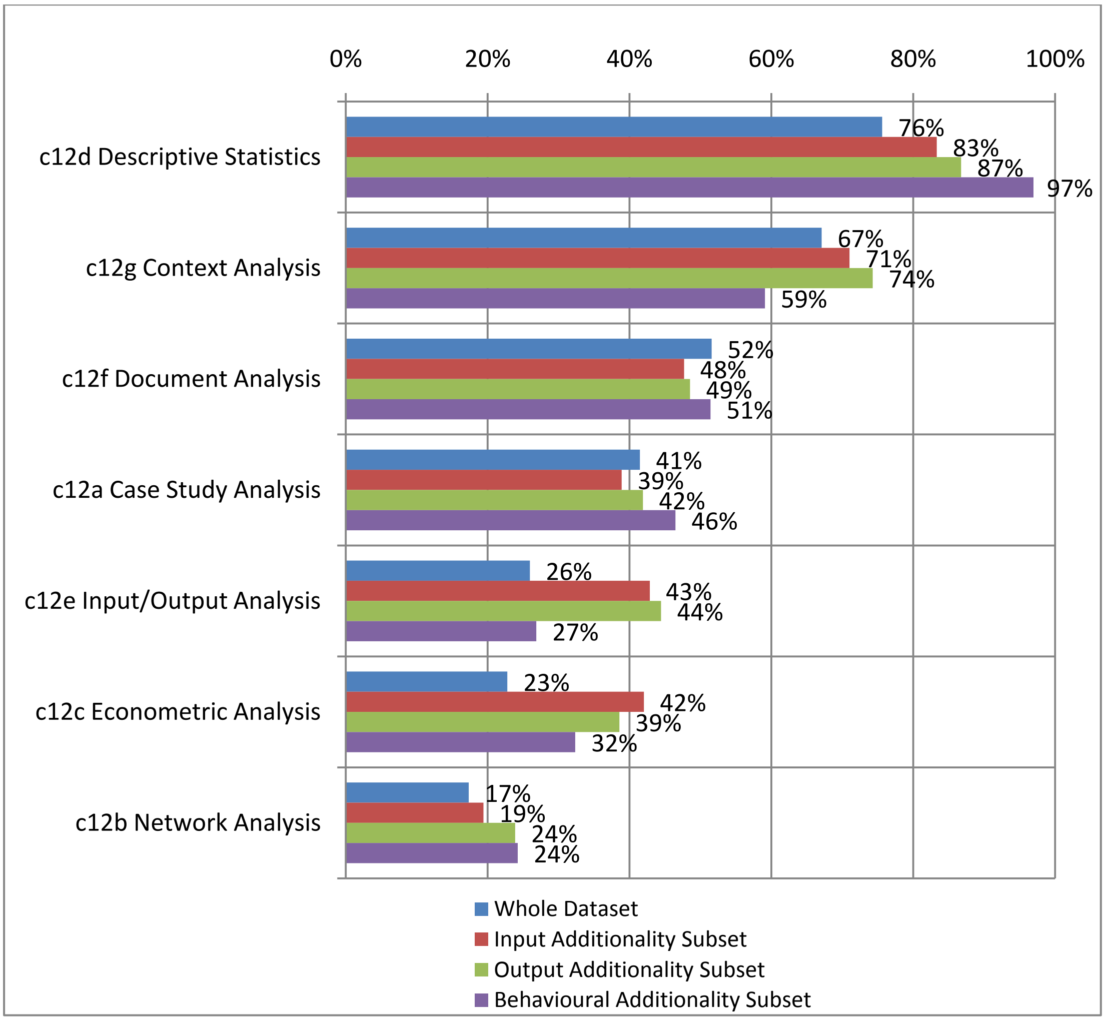
There are, however, less differences between the three types of additionality evaluations as regards econometric analysis. All three of them utilise it statistically significantly more often than the rest of the sample, even if behavioural additionality is slightly less linked to econometric analysis.

Around one-fifth of the evaluations in the whole sample used network analysis as the data analysis method. The results for all three types of additionality are statistically significantly higher, around one-fourth. This is not a surprising result especially considering the fact that behavioural additionality is generally associated more with networking.

Case study analysis was used by around 41% of the whole subset and the results for input and output additionality are similar. One of the most startling results in this Chapter is that although the ratio of evaluations using the case study analysis is slightly higher for evaluations that cover behavioural additionality (46%) than the whole data, the association between the case study analysis and behavioural additionality is very weak. This is extremely puzzling on the face of the results obtained in the previous part: behavioural additionality evaluations utilise non-experimental methods as they use interviews significantly more but at the same time their link to case study analysis is not as strong as one would expect. One would only explain this with the probability that behavioural additionality evaluations predominantly use qualitative methodologies, but as the question of behavioural additionality is quite complex and situation-dependant, the data is collected through interviews and/or surveys and/or both.

To sum up, the main data collection analysis methods for behavioural additionality were descriptive statistics (97%, significantly more than the others), context analysis (59%), document analysis (51%), case study analysis (46%), input-output analysis (27%), econometric analysis (32%, significantly more than the others) and network analysis (24%, significantly more than the others). In addition, behavioural additionality deviates in some aspects of methods from input and output additionality evaluations, the former more often employing interviews and surveys.

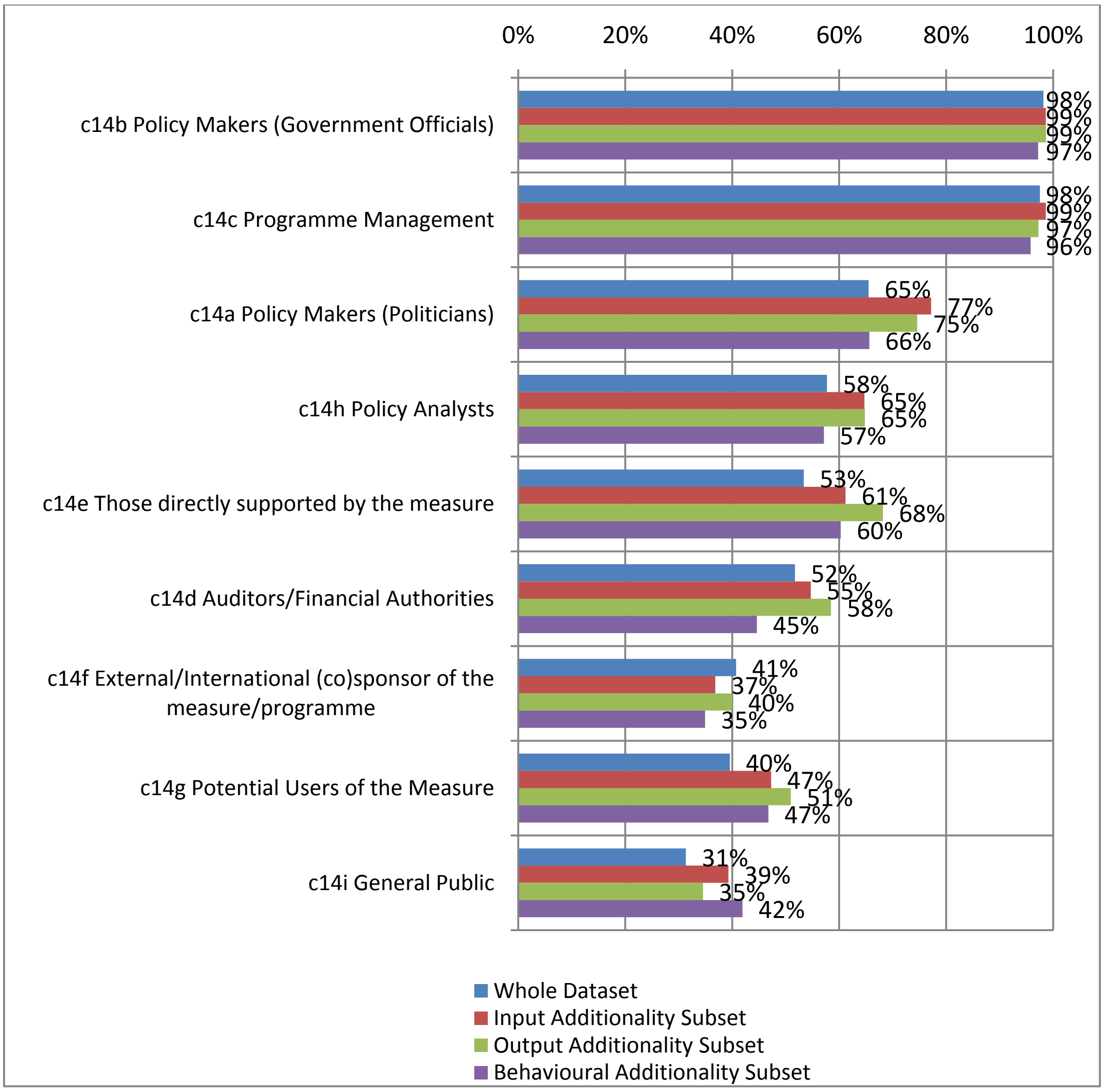
Exhibit 61: Data Analysis Methods and Additionality



3.14 Main Intended Audiences of Evaluation

The main intended audiences of evaluation in the INNO-Appraisal sample were programme management (98%), government officials (98%), politicians (65%), financial authorities (52%), policy analysts (58%), those directly supported by measures (53%), external sponsor of programmes (41%), potential users of measures (40%) and finally the general public (31%) as depicted in Exhibit 62. For input and output additionality, these ratios are statistically significantly higher for financial authorities (around 10% higher), policy analysts (around 10% higher), potential users (around 10% higher) and those directly supported by the measure (around 15% higher). The last two categories are also higher for behavioural additionality (around 10% higher). Similarly, for behavioural additionality the ratio of general public is 10% more than that of the whole set. This slight emphasis on the potential and actual users of measures is probably due the reason that the question “what difference does it make?” of additionality is most relevant to these categories of audiences. Furthermore, the fact that behavioural additionality is less associated with the auditors/financial authorities supports the previous finding that the financial and (tangible) economic implications of behavioural additionality are less obvious, and that auditors struggle with quantifying – and thus appreciating – the effects.

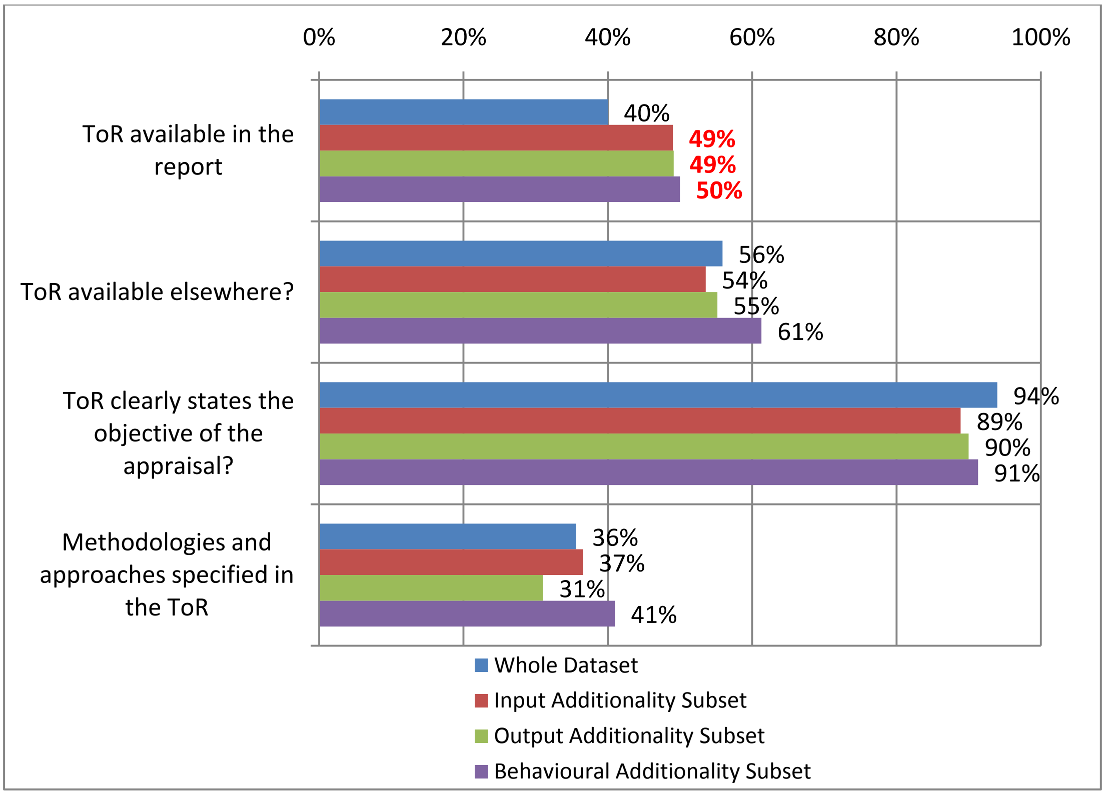
Exhibit 62: Main Intended Audiences of Evaluation



3.15 Terms of Reference

Exhibit 63 depicts that for around 40% of the evaluations, terms of reference were available. For additionality subsets this ratio is slightly higher. Among those evaluations that terms of reference were available, 94% of them clearly stated the objectives. Similarly, 36% of evaluations specified the methodologies and approaches in their terms of references. One implication here is that behavioural additionality is mostly a specified and client-driven topic of evaluation. This is important, and confirmed in one of the case studies below (case 3), however, there are cases in which evaluators push reluctant programme owners towards the concept (case 1).

Exhibit 63: Terms of References of Evaluations



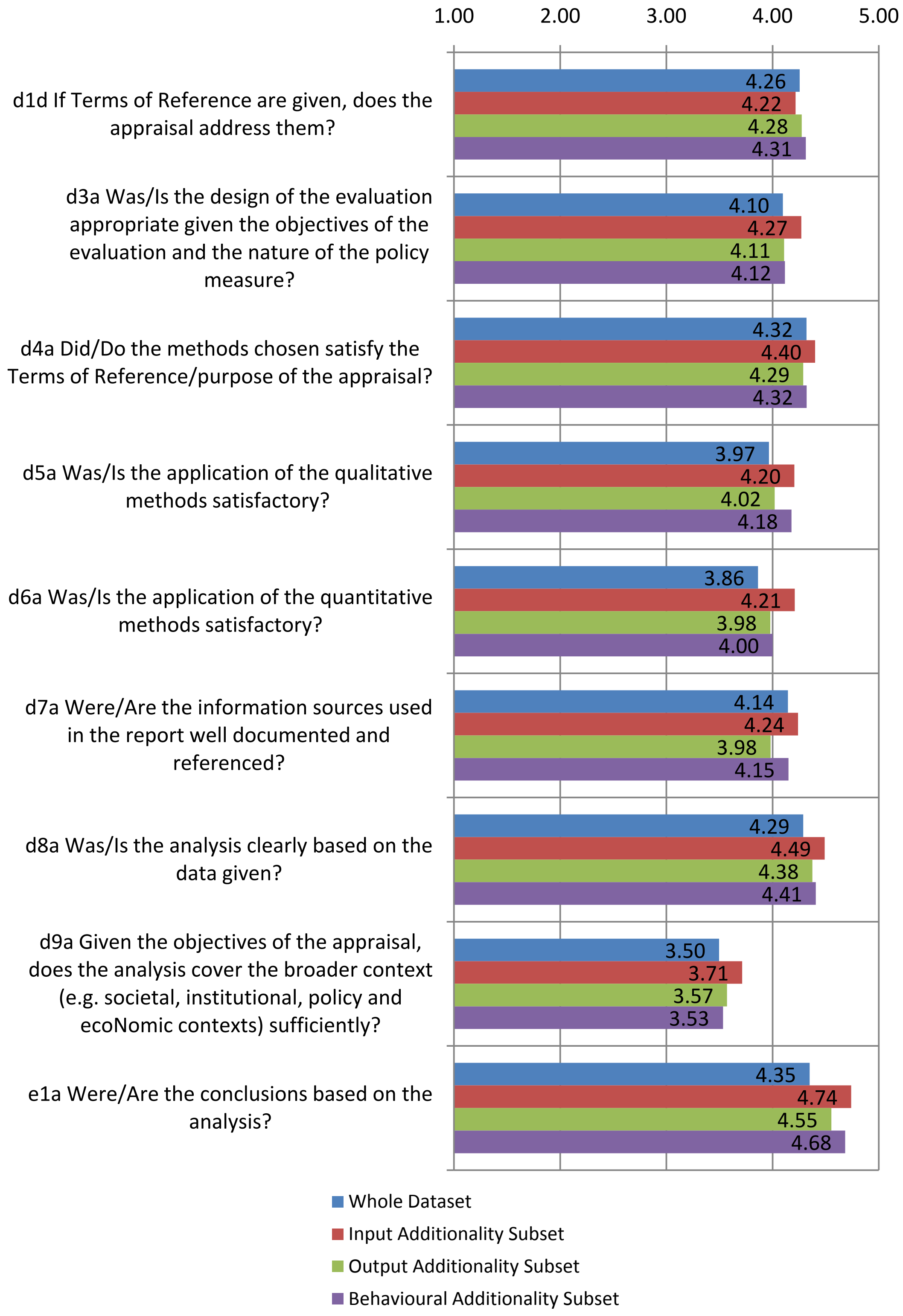
3.16 Quality of Evaluations

Several dimensions of the perceived qualities of the evaluations in the INNO-Appraisal sample are depicted in Exhibit 64. The template used a 5 points Likert scale ranging from “1, no, not at all” to “5, yes, definitely” and intermediate values were not defined. As it can be read from Table 80 and Exhibit 64, input additionality evaluations were perceived as higher quality in all dimensions including the quality of the application of the qualitative methods. Behavioural additionality evaluations were only perceived more quality in dimensions related to “the quality of the application of the qualitative and quantitative methods”, “analysis being based on the data” and “conclusions being based on the analysis”.

Table 80: Summary of the Perceived Quality of Additionality

	Input Additionality Subset	Output Additionality Subset	Behavioural Additionality Subset
d1d If Terms of Reference are given, does the appraisal address them?	No statistical difference	No statistical difference	No statistical difference
d3a Was/Is the design of the evaluation appropriate given the objectives of the evaluation and the nature of the policy measure?	higher	No statistical difference	No statistical difference
d4a Did/Do the methods chosen satisfy the Terms of Reference/purpose of the appraisal?	higher	No statistical difference	No statistical difference
d5a Was/Is the application of the qualitative methods satisfactory?	higher	No statistical difference	higher
d6a Was/Is the application of the quantitative methods satisfactory?	higher	higher	higher
d7a Were/Are the information sources used in the report well documented and referenced?	No statistical difference	No statistical difference	No statistical difference
d8a Was/Is the analysis clearly based on the data given?	higher	higher	higher
d9a Given the objectives of the appraisal, does the analysis cover the broader context (e.g. societal, institutional, policy and ecoNomic contexts) sufficiently?	higher	No statistical difference	No statistical difference
e1a Were/Are the conclusions based on the analysis?	higher	higher	higher

Exhibit 64: Quality of Evaluations and Additionality

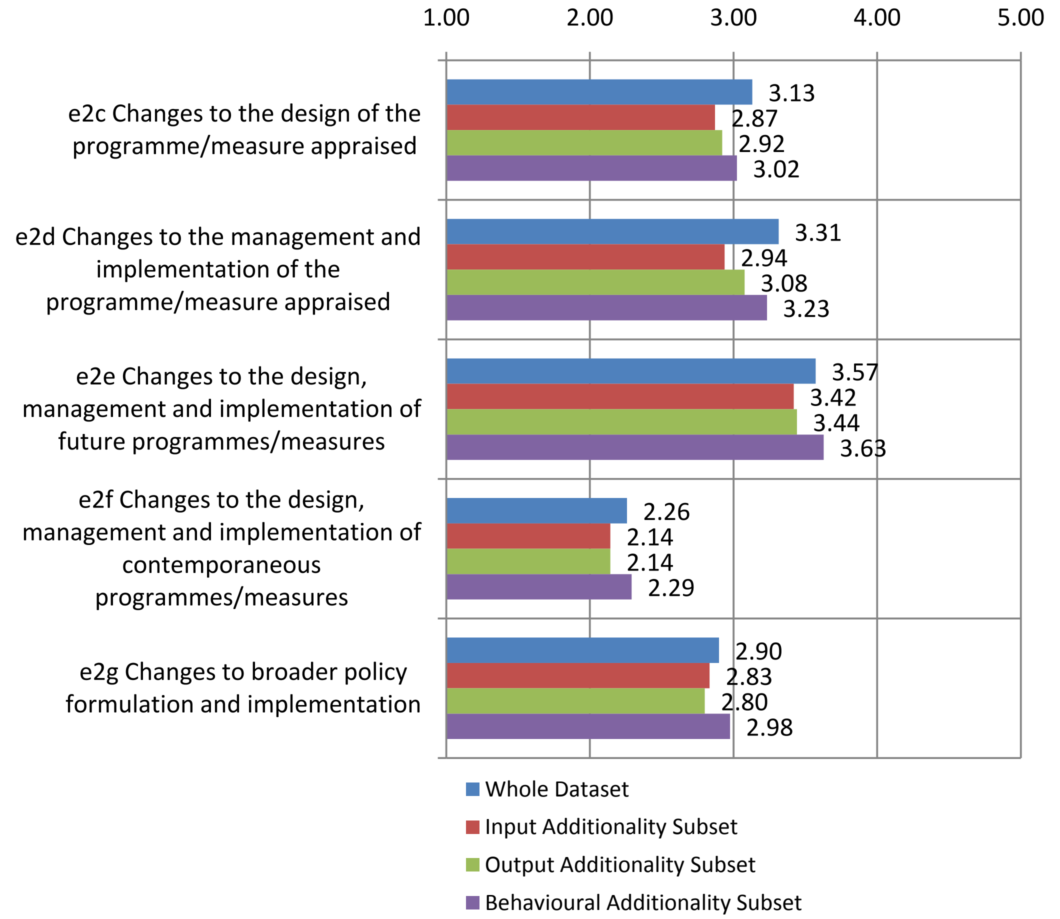


Note: Scale is 5 point Likert. "1" wass given as "No, not at all", "5" as "Yes, definetely" and intermediate values were not defined.

3.17 Usefulness of recommendations

As shown in Exhibit 65, in general the usefulness of recommendations for the evaluations was not perceived very highly compared to, for example, quality. All three subsets of additionality scored close to the general dataset, which shows no significant difference in terms of the usefulness of recommendations. Input additionality is to a small scale negatively correlated with usefulness of recommendations regarding “management and implementation of the programme design”. Behavioural additionality evaluations, despite being not as useful as the whole dataset, were slightly more useful than input and output additionality. In other words, behavioural additionality is the most useful of all three kinds of additionality evaluations, which are not in fact more useful than other evaluations.

Exhibit 65: Usefulness of Recommendations of Evaluations and Additionality

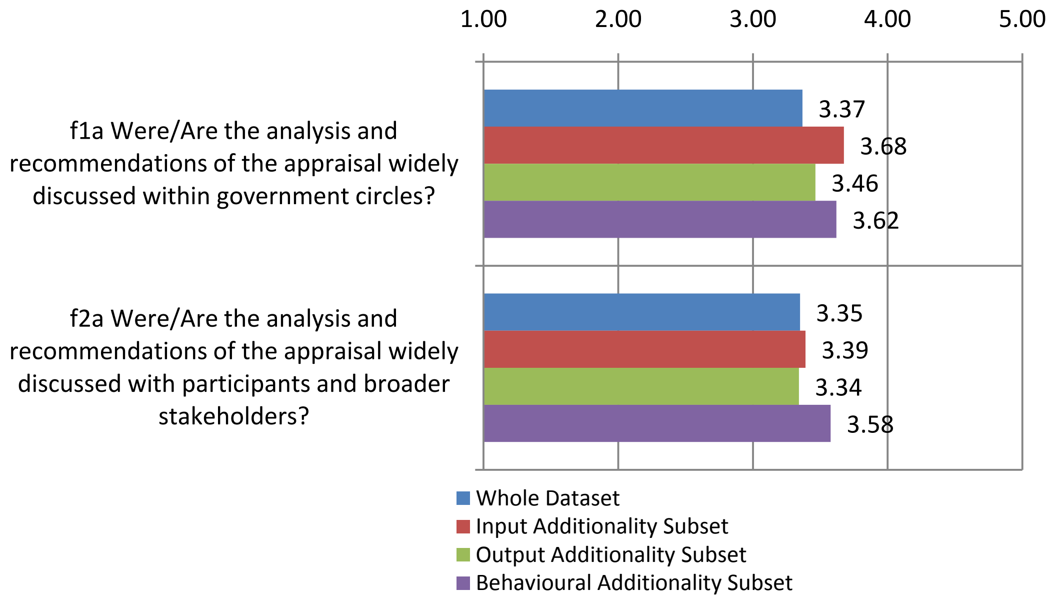


Note: Scale is 5 point Likert. "1" wass given as "No, not at all", "5" as "Yes, definetely" and intermediate values were not defined.

3.18 Consequences of Evaluations

As shown in Exhibit 66, the scope of the evaluations’ discussion within government circles scored 3.37 in the above used 5-point Likert scale. Similarly, discussions with stakeholders scored 3.35. These results imply that it is perceived that considerable attention to evaluation is given by the government officials and wider stakeholders which, however, seems to have room for improvement. Evaluations that cover behavioural additionality are discussed more widely than the whole dataset, both within government and – more pronounced even – with stakeholders more widely, which is consistent with the finding that behavioural additionality is statistically significantly associated with audiences such as general public and beneficiaries. Interestingly, output additionality is not discussed more broadly within or outside government circles, contrary to the general belief that discussion on innovation policy focuses mainly on output.

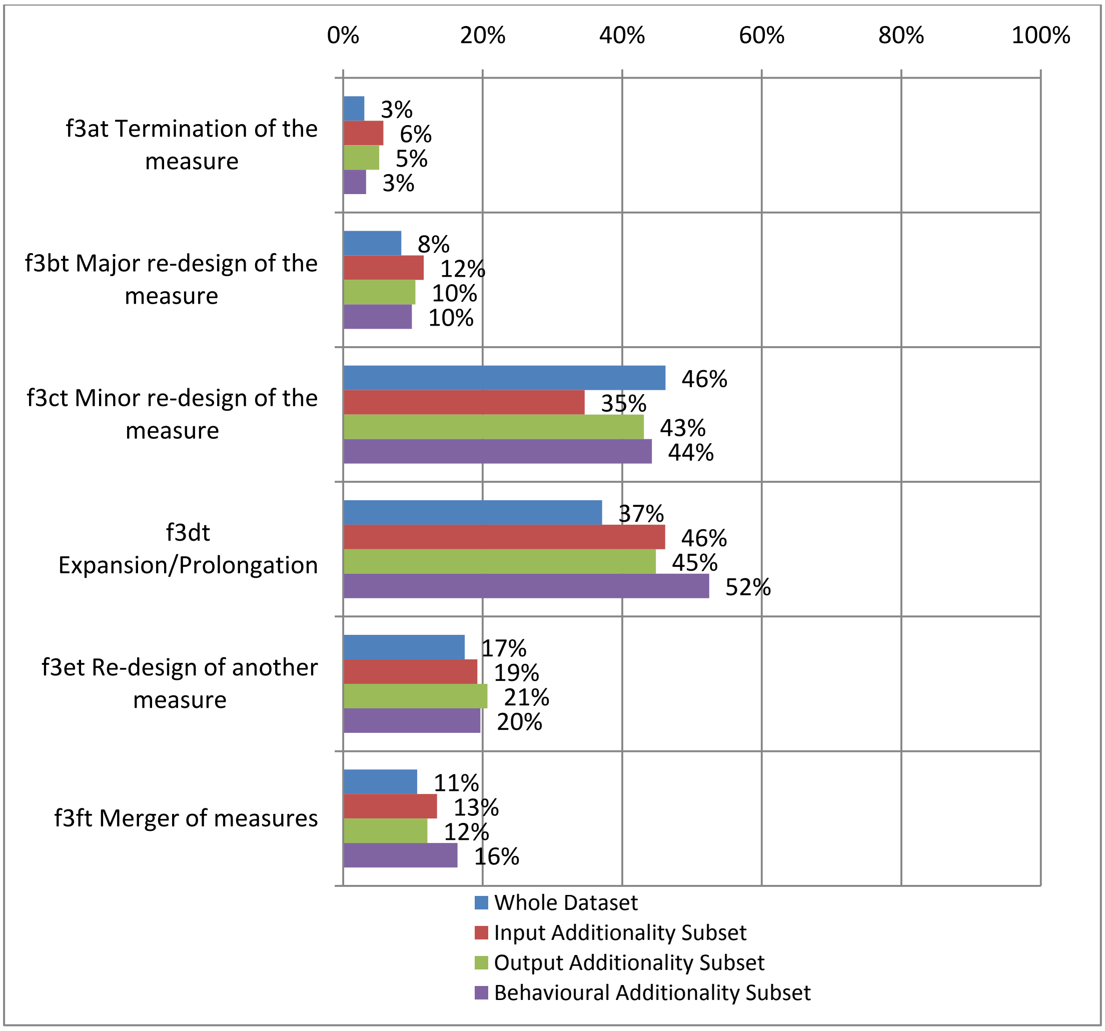
Exhibit 66: Discussion of the Results and Additionality



Note: Scale is 5 point Likert. "1" wass given as "No, not at all", "5" as "Yes, definetely" and intermediate values were not defined.

Exhibit 67 shows the consequence of evaluations as they were addressed in the template. Within the INNO-Appraisal sample, only a tiny fraction of evaluations led to the termination of measures (circa 3%). The share of measures that having undergone a minor and major re-design because of evaluation is 46% and 8% respectively. Input and output additionality subsets tend to result in statistically significantly less minor re-design than the whole dataset. The percentages for "expansion / prolongation of the measure", "re-design of another measure" and "merger of measures" are 37%, 17% and 11% respectively for the whole sample. For both prolongation/expansion and re-design of another measure, all three additionality subsets yield statistically higher ratios. As already said in Chapter 3.3, behavioural additionality's strong association with the expansion/prolongation of the related measure could indicate that the concept helps policy makers to understand that time is needed for behavioural changes to show effects at the innovation end, and – in addition – that it is used for legimisation.

Exhibit 67: Consequences of Evaluations and Additionality



4 Text Analysis

As outlined in section 1 of this chapter and in Gok (2010), there are a number of different and sometimes conflicting beliefs relating to behavioural additionality in scholarly literature (see Table 79 above for a summary). The four broad categories of definitions of behavioural additionality are as following:

A. a simple extension to input and output additionality,

B. the change in the non-persistent behaviour related to R&D and innovation activities

C. the change in the persistent behaviour related to R&D and innovation activities

D. the change in the general conduct of the firm with substantial reference to the building blocks of behaviour

This merits further investigation in our empirical dataset on evaluation practice for at least two reasons. First, it is very important to see how behavioural additionality is really understood in an applied real-life context, to improve its application, and to arrive at a generally shared understanding. Secondly, to be able to develop a new theoretical/conceptual framework, it is imperative to understand how discussions in scholarly literature influence the perceptions in practice.

Further, the text analysis was also the basis for the selection of cases for a more in-depth, interview based analysis of the use and usefulness of BA (see next section).

Carrying the analysis presented in detail in Gok (2010) further, the evaluation reports in the sample covering behavioural additionality are analysed as to their understanding of the concept. 33 reports out of 81 were looked at for the definition and usage of the concept of behavioural additionality. The selection of the sample was based on the necessary threshold of quality of the evaluation as well as on pragmatic consideration (language skills in the team). The number 33 is sufficient for a case by case analysis of different BA understanding, as this analysis is not about being representative, but explorative. It should also be noted that not all of these 33 reports mentioned the term behavioural additionality explicitly but in all of them the idea of the concept in one way or another was present.

The text analysis of the 33 reports covering behavioural additionality investigated the following important dimensions that were featured in the analysis of scholarly literature conducted in the previous chapter:

- Was behavioural additionality explicitly mentioned or used implicitly?
- what was the definition of behavioural additionality?
- did the implicit or explicit definition of behavioural additionality include the elements of persistency (as persistency is at the core of behavioural additionality as a consequence of learning and change processes)?
- was the implicit or explicit definition of behavioural additionality confined to R&D behaviour only or to collaboration behaviour only?
- were there are any references to the individual building blocks of behaviour (or was the concept not differentiated into such building blocks)?

The result of the analysis is presented in Table 81 below. It reveals that the typology of the literature is replicated in evaluation practice, with a few minor differences. First of all, the two distinct categories of approaches to behavioural additionality such as "the extension for input and output additionality" (category A) and the "change in the non-persistent behaviour related to R&D and innovation activities" (category B) form a single category in the practice (category A+B) as it was not possible to distinguish category A from Category B in most of the cases. This means that there is one category of understanding that does not differentiate, that sees behavioural additionality as a residual category, all kinds of changes not attributed to input or output additionality are labelled as behavioural additionality, with no linkages to persistency of that change.

In contrast, the category C in the literature ("the change in the persistent behaviour related to R&D and innovation activities") is slightly more differentiated in practice and thus can be split into two categories, those that look at behavioural additionality as one distinct phenomenon (Category C1), and those that differentiate it into various building blocks (category C2) is created. This means that there are evaluations that look at persistent change in R&D and innovation related behaviour, but do not differentiate, do not dig deeper as to what the different elements are that characterise that change, and there are those that do the latter, but still do not enlarge the concept to the conduct of the firm more broadly (which would be category D).

The analysis showed that those four categories are more or less evenly distributed among the 33 evaluations. There is, in other words, no dominant understanding of behavioural additionality. While some evaluators define it extremely narrow, or not at all (as a residual category), others have a differentiated concept that defines and operationalises building blocks and looks beyond R&D and innovation activities. Furthermore, there is a clear link between the scope of the behaviour they investigate and the definition category. For instance, while category A+B evaluations are mostly limited to collaboration behaviour, Category D evaluations have a much wider scope.

Table 81: Classification of Definitions of Behavioural Additionality in the Practice

	Category A + B	Category C1	Category C2	Category D	Other
Definition	The non-persistent extension of input additionality OR the change in the non-persistent behaviour	change in the persistent behaviour	change in the persistent behaviour with minor references to building blocks	change in the general conduct of the firm, reference to building blocks of behaviour	Inconsistent OR not possible to analyse
Coverage	Only R&D and innovation	Only R&D and innovation	Only R&D and innovation	(Some of Them) Beyond R&D and innovation	
Persistence	One-off, no persistence OR Rather mid-term than long-term and rather less persistent	Persistent OR Rather long-term than short-term and rather more persistent	Persistent	Persistent	
Observations	30% of evaluations 30% only collaboration BA mostly implicit	15% of the evaluations 50% only collaboration BA mostly implicit	20% of the evaluations Mostly collaboration and beyond BA mostly implicit	15% of the evaluations Mostly collaboration and beyond BA is more explicit than other	20% of the evaluations

5 Case Studies

5.1 Introduction

The previous sections have shown that the application of behavioural additionality is very common practice in innovation policy evaluations. Evaluations applying it have a set of distinct features, but the concept – although powerful as a learning and assessment tool – is still not clearly defined, has a lot of implicit or explicit variants, is still poorly operationalised in those different variants and thus appears to be underexploited.

Therefore, this last section summarises the results of a set of completed case studies done. These case studies analyse the application of behavioural additionality in a few carefully selected evaluations. The objective of this analysis is to better understand:

- how the concept is operationalised and applied and what methods are used in evaluations that appear to be of high quality (good case approach)
- how context conditions and properties of innovation policy measures shape the application of the concept
- what challenges the concept faces in its application and in the exploitation of results
- what specific issues arise in the interaction between evaluators and responsible policy makers when such a complex concept is applied
- how the results of behavioural additionality are used

The selection of the cases built upon the text analysis discussed in the previous section. The basic idea of these cases is to learn from good applications. Thus, out of the 33 cases included in the text analysis, those cases were selected that applied the concept in the broadest and most thorough way. The first selection filter was: did the concept play a prominent role in the evaluation, implicitly or explicitly. The second criterion was: was the application of the concept thorough and promising to yield some meaningful insights? A third filter then was to make sure different variants of the concept were included, ranging from a simple understanding of behavioural additionality as an increase in collaboration (which then, within this limitation was conducted carefully and in a sensible way) to complex differentiations of various building blocks of behaviour. And finally, the selection made sure that different kinds of policy measures were covered.

This analysis took into account the country contexts within which the cases were performed, but it did not – could not – strive to cover different countries and their evaluation context; given the heterogeneity across Europe this would not have been pragmatic. The case analysis builds on a thorough investigation of the evaluation reports and the logic of the underlying policy measure, an analysis of the respective template that was filled in by policy makers, and then subsequent telephone interviews with the lead evaluators and – in most cases – the responsible policy maker. The semi-structured interviews followed a basic interview template (see annex 6A). Table 82 gives a list of the cases.

Table 82: The Cases of behavioural additionality

Case Number	Case Title
Case 1	Evaluation of SME innovation cooperation measure: evaluator driving the application of behavioural additionality
Case 2	Evaluation of a measure to improve capabilities and effectiveness of innovation support agencies and the effect of the improvement on firms
Case 3	Evaluation of a measure to stimulate the foundation of firms in biotech sector, mobilise more risk funding and improve networking to turn the country into a leading edge biotech location
Case 4	Evaluation of a programme that intends to create three-way partnerships between businesses, recently graduated people and senior academics acting as supervisors
Case 5	Evaluation of a classical continental regional grant programme run by a dedicated agency

The following section summarises the major lessons from the cases. There is not enough space and no need to give a full account of each case; we will concentrate on the main features and lessons in each case.

5.2 The cases

Case 1: Evaluation of SME innovation cooperation measure: evaluator driving the application of behavioural additionality

The first case is chosen because it shows the dynamic and contested nature of complex behavioural additionality evaluations and the merits and challenges for policy makers and evaluators. It is an example of an innovation policy measure that sought to improve and broaden the cooperation of SMEs with each other and, more importantly, with Universities and other research institutes. It built upon earlier schemes of supporting collaboration that resemble institutional funding for public research organisations that engage in contracts with individual firms and was limited and inflexible in its outreach. The new scheme wanted to create broader networks and wanted to support projects that would then lead to market introduction. The change in behaviour it intended was thus on two levels: (1) on the firms and the public research partners to build new forms of collaborations, gain new innovation and management capabilities and follow them up through market introduction and (2) on the policy community more broadly to change the way cooperation programmes are designed for SMEs. The responsible programme officers were dedicated not only to the programme, but to the evaluation, which they perceived as a learning and justification tool for the change in their approach.

The evaluation that was commissioned was a mix of summative and formative. However, it wanted the evaluators to focus on the effectiveness in terms of market introduction of results stemming from the support, neglecting to a large extent the first dimensions, creating new networks and enabling partners to engage in persistence collaboration. The evaluators suggested, during the evaluation, to enlarge the remit and include the behavioural additionality aspect, based on the insight that this in itself would be an important effect on policy, for the programme and beyond, and it would yield more immediate and clearly attributable results, while market introduction took usually a number of years and was dependent on a large set of contexts and firm specific factors, thereby reducing the possibility to attribute market success or failure to the project. During the process policy makers accepted this change of remit. The evaluation thus tackled behavioural additionality, but mainly in two aspects: new forms of persistent collaboration and the uptake of new methods and techniques in firms. The evaluation found a whole range of learning effects both directly after the finalisation of projects and a couple of years later, in a follow up analysis: enlarging of technological competence, entry into new technologies, improvement in the ability to cooperate, ability to design and produce new products and services. The policy makers were very open to learning about behavioural additionality and welcomed, after first hesitations, the change. However, they still insisted on proof of market success, not regarding the learning effects observed as sufficient to fully appreciate the programme. The methods applied to tackle the subject were a context analysis, participant survey, participation in reviews and focus groups, workshops with programme owners, and, most importantly, intensive interviews within a case study approach over a whole range of years, until a couple of years after the completion of the original project. While the dimensions of behavioural change were collaboration capabilities and attitudes, management of projects (including financial management) and technological as well as absorptive capability, the evaluation could only tackle cooperation. The real time case studies and the follow up phase after two years were not originally foreseen, but then deemed necessary given the enlarged remit. The approach, although thorough in its application and analysing persistence, could not dig deeper into the many facets of behavioural change, as a more nuanced and inter-disciplinary approach would have been needed. The evaluation itself was discussed positively and demonstrated learning effects, but programme owners had hoped for more tangible market results. The programme was stopped because of procedural issues in the management.

Main lessons: behavioural additionality can be used to understand innovation policy effects more thoroughly if it seeks to define the dimensions of behaviour affected, show if those effects are persistent and demonstrate how they lead to the intended second level effect (more innovation). Policy makers and evaluators may have different concepts of behavioural additionality, and strong communication between them is indispensible during the process. To fully grasp behavioural additionality it would have to be built in from the very beginning and an inter-disciplinary, ethnographic approach would be needed for the different facets of behaviour. This, in turn, is unlikely to be pragmatic for regular evaluations. However, it would be advisable to develop a set of experimental designs in order to understand some basic principles of behavioural change induced through policy interventions, and to derive from simplified, applicable yet meaningful evaluation approaches. Policy makers, even if open to the concept, still – at the very end – often fall back on the secondary effect mainly and to undervalue the meaning of persistent behavioural change. Even if convinced about behavioural mechanisms, for policy makers and programme owners they remain a challenge because of the time they take to show effects and because of the attribution problem.

Case 2: Evaluation of a measure to improve capabilities and effectiveness of innovation support agencies and the effect of the improvement on firms

The second case is chosen because of its methodological approach. It is the evaluation of a programme that seeks to improve the effectiveness of intermediaries who support mainly SME in innovation and innovation management. The changes induced by the programme are on two levels: Improvement of the performance of the centres in their support function to the firms, as well as more visibility; and, on a second level, improvement of the firm's innovation activities (cooperation, new networks, qualification of staff). The supported centres set up distinct projects to help firms in their innovation activities. The evaluation was mandatory, some administrators, especially those on the ground, in contact with the beneficiaries, were more in favour of a thorough evaluation of behavioural change as part of the evaluations than others.

The evaluation was entirely free in its design and took advantage of monitoring data and interviews. The leading edge of the design was in the interactive definition of the components of action and behaviour. The interviews set up a categorisation of activities and changes with the centres that then subsequently were monitored. The definition of the action dimensions was a result of deduction – from initial hypotheses – and induction – the interviews. The challenge in this approach is to objectivise the data that is largely interview based and to link the activity and the change to the final outcome (improvement of the supported firm). A two step approach was developed, from a large list of action and changes was clustered into a limited set of key parameters (on the basis of triangulation) and those parameters were then conceptually linked to effects of the programme (better support, more innovation etc.). Most importantly, in a final step those links of parameters and effects were subsequently verified in workshops with centre staff. If established, the change in the action parameter (which is traced through monitoring interviews) is taken as effect of the programme on the first level and at the same time is used as a proxy for change at the second level. Thus, this methodology forgoes quantitative measurement of change in the second level and ex post attribution of action change and effects. Interestingly, the concept of behavioural additionality was never undefined, but the evaluation was about a change of cooperation behaviour and skills of firms (second level) and quality of the centres (first level), and the basic idea was to measure the sustainability of that change.

Major Lesson: it is possible to differentiate behavioural additionality and define the building blocks of behaviour as well as chain of effects. This can be done in a mix of deductive and inductive approaches, with a focus on interaction with the beneficiaries. A concept of measuring action and change can be developed to link with the secondary effects of the intervention (the final improvements within firms). The process of defining the concept is in itself a learning tool for all involved, as it forces to think consciously about changes and their effects. The monitoring can be optimised and simplified. However, the change - effect linkages need constant supervision and a through triangulation with other parties and through group approaches. Moreover, a cross check with final outcomes on the side of end beneficiaries is asked for, not relying entirely on the proxy indicator action change in the agencies (the intermediaries). Such an elaborated approach is also easier to implement with intermediaries whose mission it is to improve the system (i.e. innovation agencies) rather than end beneficiaries for which time commitment for the approach may be challenging.

Case 3: Evaluation of a measure to stimulate the foundation of firms in the biotech sector, mobilise risk funding and improve networking to create international attraction

This case is selected because of the elaborated methodology that was initiated by policy makers rather than evaluators, inducing a joint learning about and implementation of the behavioural change concept. The change which the measure wants to induce, as the title of this case suggests, is multi-dimensional, a general change of attitude and practices in the biotech sector, involving risk taking (capital, firm creation), capability build up in firms, clustering and build up of actor capability of clusters and regions. The evaluation was summative and formative, as one of the first large cluster programmes of the evaluation was intended to make the programmes improve and to market them for further applications. The design of the evaluation was entirely free; next to an initial workshop with a commissioning agency there was not much interaction. Interestingly, the programme owner insisted on a qualitative part of the evaluation, interviews, understanding the complexity of the programme cannot be grasped with survey and monitoring data alone. Behavioural additionality was not explicitly defined, but measured through a control group approach (firms funded, firms not funded). This was tackled with a combination of three instruments, and only this combination is seen as having yielded the necessary depth and breadth of results: Analysing existing economic data on firm foundations, turnover and collaboration, getting more tailored data through surveys and following up those surveys with in-depth interviews, focusing on their risk taking and capability development. The latter was not planned in the initial design, but deemed important to capture the nature of the changes. The ministry responsible for the programme was content with the results of the programme demonstrated by the evaluation. Their main interest was in the long term effect of capabilities, new capacities (large number of establishments) and networked clusters in the regions, and this was demonstrated. The evaluation was used to market the approach and the successful regions, with an entirely independent evaluation approach.

Lessons: As in earlier cases, this case has well demonstrated the need for qualitative, in-depth interviews and cases with follow up over time to complement survey and other statistical data. The full effect of behavioural additionality evaluation needs this mix. It also showed that policy makers interested in the long term behavioural effect of a programme can influence evaluation (the evaluators wanted quantitative evaluation only) without unduly interfering with the analysis. Early communication about expectations was key, not so much the text as the Terms of References.

Case 4: Evaluation of a programme that intends to create three-way partnerships between businesses, recently graduated people and senior academics acting as supervisors

The case is selected because the programme objectives are very relevant to the idea of behavioural additionality. This long-running programme helps firms to conduct projects in which they hire young people who would work in collaboration with an academic from a university. This facilitates a university-industry link and creates new firm capabilities, contributes the training of young people and ultimately influences company behaviour through the partnership. Therefore, in principle the programme intends to change the behaviour of the firms, the young person who is at the beginning of their career and the university through the academics who act as supervisor. The extent of behaviour in focus here is very wide and persistency has always been a concern in the change of behaviour. The measure belongs to a country where evaluation culture is very advanced in all levels of policy and society and the idea of additionality is very well embraced.

The evaluation was designed as summative. The main objectives of the evaluation were to investigate the uptake, economic impacts and cost effectiveness of the programme. Additionality, in the form of input and output, was the main dimension of the evaluation. The evaluation focused on questions such as how much extra revenue firms created because of the programme or how many extra jobs were created.

The evaluation was designed mainly by the programme management and the policy-maker had little to say. The main purpose of the evaluation was legitimisation apart from the compulsion to conduct an evaluation in a given interval. Behavioural additionality was only considered as an extension of input and output additionality. The programme manager felt that the questions of input and output additionality were too limiting and needed to be extended to be able to reveal the full scale of the effect that the programme was to create. For example, they asked the evaluator to investigate the scale and the scope of the projects if there was no support from the programme, as well the classical additionality question of whether they would have conducted the project had there been no support. These questions were asked only as part of input and output additionality in a survey setting and have never been carried further. Similarly, operational learning has never been an issue in the evaluation and therefore behavioural additionality is not considered as part of this. Finally, the evaluator did not agree with the idea of investigating behavioural additionality and did not provide any explanations or recommendations for the basis of the limited coverage of the concept in the evaluation. The programme manager indicates that they now see the importance of behavioural additionality and try to include this dimension in further evaluations.

Lessons: This case presents two main lessons. First, even for measures which consider behavioural change as their ultimate objective, evaluation of behavioural additionality is not always relevant. Political factors, especially the need for legitimacy, form the main reason for evaluation. The full benefits of behavioural additionality is not utilised mainly because they are considered as irrelevant or unnecessary. This suggests that the relationship between the rather evolutionary rationale of some programmes (changing behaviour, elevating cognitive capacity as discussed in Section 2 of this chapter) and the evolutionary concept of behavioural additionality does not always link together, especially in the mind of the policy-makers and programme managers. Secondly, this case clearly shows a move towards better appreciation of what behavioural additionality can offer.

Case 5: Evaluation of a classical continental regional grant programme run by a dedicated agency

The reason this case is selected is because of its elaborated methodology. Furthermore, this programme was one of the first programmes that evaluated behavioural additionality in subsequent evaluations. The programme is a classical, continental grant programme, in which firms apply to a regional agency for R&D funding. The programme has very few priorities and does not have any objectives in terms of influencing the supported firms' behaviour. The main performance criterion for the agency is increase in total business R&D spending by creating input additionality.

The evaluation focused exclusively on behavioural additionality with a view that other evaluations do not create enough legitimacy for the programme which needs political support against the demands of the policy-maker who fund the agency. The evaluation was conducted by a scholar who has been working with the agency on different projects for a long time and who has almost been considered as a member of the staff. The head of the agency directly asked the evaluator to conduct an evaluation of the effects that would reveal the true benefit of the programme so that the comparatively low degree of input and output additionality can be defended. The evaluator worked almost in isolation and did not interact considerably both with the evaluation unit and the programme management within the agency. Due to the fact that the evaluation was considered as a tool for providing backing and legitimacy to a programme that was in question, the methodology was designed to be very elaborate to yield robust results. This evaluation used the most advanced behavioural additionality evaluation design with multiple control-groups, counter factual approach, qualitative and quantitative data collection methods and several phases of piloting. The evaluation used the rather simple but the most pragmatic definition of behavioural additionality (Falk, 2006, 2007).

The evaluation did not create any recommendations, this was not intended from the very beginning. Similarly, as the evaluator worked directly with the top management of the agency and interacted little with the rest of the agency, the "process benefit" of the evaluation in terms of operation learning was limited. The evaluator thinks that they should have included the secondary effects of the programme on the firms that were not supported but working with the supported firms, in a view that these effects could have further supported the conclusion of the evaluation. The evaluation provided a solid ground for the continuation of the funding for the programme but nothing has changed in the agency after this evaluation in terms of its operation.

Lessons: This case shows the variety and sophistication in behavioural additionality approaches. However, it also demonstrates that the sophistication of the method alone is not sufficient to yield optimum results out of the evaluation on the operational level within programme management. While the evaluation demonstrated behavioural additionality effects, it did not provide suggestions for consequences out of those findings, and thus did not provide guidance for the follow up. The evaluation was thus almost over-designed, putting too much emphasis on showing effects and not enough on embedding the process within the agency. However, the process of evaluation for behavioural additionality is important. Unlike some other concepts of evaluation, the idea of behavioural additionality focuses on the process of the programme as the interaction between the firms and the agency, only their involvement can ensure the full utilisation of the process benefits that behavioural additionality can offer.

6 Conclusion

Since evaluations in innovation (and science) policy have introduced the concept of behavioural additionality, it has stirred controversy as to its specific character, its preconditions and its usefulness. Very often the concept appears to be applied but is ill-conceived or only used in a very limited understanding. <ethods are not appropriate, and the multiple dimensions of behaviour and the cascade effects of changes in behaviour on innovation performance and management more generally are not conceptualised.

There is a consensus in academic literature that the concept is crucial, but there is no consensus in the literature about what behavioural additionality exactly is and means. While some scholars put it at the very heart of the evolutionary structuralist understanding for innovation policy, others see it as an additional dimension that may fill gaps in understanding policy effects.

Empirically, the text analysis and the cases have shown that with small adjustments, the confusion over the definition of the concept in scholarly literature prevails in the field of practice as well. Furthermore, none of the four different understandings are dominant. We also see that additionality has different evaluation characteristics than evaluations covering other topics.

As the effects of behavioural additionality on innovation dynamics in firms are complex, time consuming and intertwined with other influences, evaluations must clearly demonstrate the conceptual link between behavioural change and the innovation effect. They then must empirically grasp the change in behaviour and try to find robust indications that the link to innovation effects exists. To enable such simplification, the concept still needs a better theoretical foundation to derive a more suitable framework of analysis with a better defined unit of analysis – behaviour and its various components (Gok, 2010).

Furthermore, the concept needs methodological clarifications. Empirically, behavioural additionality evaluations are quite different in many aspects of methodology and use of the evaluation than input or output additionality. However, we also observe a broad diversity of methods used. There is no clearly defined set of methods dominant in evaluations of behavioural additionality, and even the most sophisticated approaches fail to ask the “right questions”. As a comprehensive understand of behavioural change would need a broad mix of methodologies that cannot be applied in all cases, experimental methodological developments are required that allow a simplified yet relevant set of approaches.

The case studies have shown that behavioural additionality is used for very different objectives ranging from legitimisation to operational learning. The complexity of behavioural additionality asks for a strong interaction and communication between those commissioning the evaluation and the evaluators. For operational learning and sound policy feedback, key concepts as to the link of behaviour changes to innovation must be shared and expectations clarified early on. Sophisticated methods alone do not guarantee the full benefits of the concept, their applications and the results must be intensively discussed among all stakeholders involved.

References

Aslesen, H. W., Broch, M., Koch, P. M. and Solum, N. H. (2001). *User Oriented R&D in the Research Council of Norway*. Technopolis RCN Evaluation. Oslo.

Bach, L. and Matt, M. (2002). Rationale for Science & Technology Policy. *In:* Georghiou, L., Rigby, J. & Cameron, H. (eds.) *Assessing the Socio-Economic Impacts of the Framework Programme (ASIF).* Report to European Commission DG Research.

Bach, L. and Matt, M. (2005). From Economic Foundations to S&T Policy Tools: a Comparative Analysis of the Dominant Paradigms. *In:* Llerena, P. & Matt, M. (eds.) *Innovation Policy in a Knowledge-Based Economy: Theory and Practice.* Berlin Heidelberg: Springer.

Buisseret, T. J., Cameron, H. M. and Georghiou, L. (1995). What Difference Does It Make - Additionality in The Public Support Of R&D In Large Firms. *International Journal of Technology Management,* 10**,** 587-600.

Busom, I. and Fernandez-Ribas, A. (2008). The impact of firm participation in R&D programmes on R&D partnerships. *Research Policy,* 37**,** 240-257.

Clarysse, B., Bilsen, V. and Steurs, G. (2006). Behavioural Additionality of the R&D Subsidies Programme of IWT-Flanders (Belgium). *In:* OECD (ed.) *Government R&D Funding and Company Behaviour: Measuring Behavioural Additionality.* Paris: OECD.

Davenport, S., Grimes, C. and Davies, J. (1998). Research collaboration and behavioural additionality: A New Zealand case study. *Technology Analysis & Strategic Management,* 10**,** 55-67.

Falk, R. (2006). Behavioural Additionality of Austria's Industrial Research Promotion Fund (FFF). *In:* OECD (ed.) *Government R&D Funding and Company Behaviour: Measuring Behavioural Additionality.* Paris: OECD.

Falk, R. (2007). Measuring the effects of public support schemes on firms' innovation activities: Survey evidence from Austria. *Research Policy,* 36**,** 665–679.

Fier, A., Aschhoff, B. and Löhlein, H. (2006). Behavioural Additionality of Public R&D Funding in Germany. *In:* OECD (ed.) *Government R&D Funding and Company Behaviour: Measuring Behavioural Additionality.* Paris: OECD.

Georghiou, L. (1998). Issues in the Evaluation of Innovation and Technology Policy. *Evaluation,* 4**,** 37-51.

Georghiou, L. (2002a). Impact and Additionality of Innovation Policy. *IWT-Studies,* 40**,** 57-65.

Georghiou, L. (2002b). Innovation Policy and Sustainable Development: Can Innovation Incentives make a Difference? *IWT-Studies***,** 57-65.

Georghiou, L. (2004). Evaluation of Behavioural Additionality. Concept Paper. *IWT-Studies,* 48**,** 7-22.

Georghiou, L. (2007). What lies beneath: Avoiding the risk of under-evaluation. *Science and Public Policy,* 34**,** 743-752.

Georghiou, L. and Clarysse, B. (2006). Introduction and Synthesis. *In:* OECD (ed.) *Government R&D Funding and Company Behaviour: Measuring Behavioural Additionality.* Paris: OECD Publishing.

Georghiou, L. and Keenan, M. (2006). Evaluation of national foresight activities: Assessing rationale, process and impact. *Technological Forecasting and Social Change,* 73**,** 761-777.

Gok, A. (2010). *The Evolutionary Approach to Innovation Policy Evaluation: Behavioural Additionality and Organisational Routines.* PhD, The University of Manchester.

Hall, B. H. and Maffioli, A. (2008). Evaluating the impact of technology development funds in emerging economies: Evidence from Latin America. *European Journal of Development Research,* 20**,** 172-198.

Hsu, F. M., Horng, D. J. and Hsueh, C. C. (2009). The effect of government-sponsored R&D programmes on additionality in recipient firms in Taiwan. *Technovation,* 29**,** 204-217.

Hsu, F. M. and Hsueh, C. C. (2009). Measuring relative efficiency of government-sponsored R&D projects: A three-stage approach. *Evaluation and Program Planning,* 32**,** 178-186.

Hyvarinen, J. and Rautiainen, A. M. (2007). Measuring additionally and systemic impacts of public research and development funding - The case of TEKES, Finland. *Research Evaluation,* 16**,** 205-215.

Licht, G. (2003). The Role of Additionality in Evaluation of Public R&D Programmes. *11th TAFTIE Seminar on Additionality.* Vienna.

Lipsey, R. G. (2002). Some Implications of Endogenous Technological Change for Technology Policies in Developing Countries. *Economics of Innovation and New Technology,* 11**,** 321 - 351.

Lipsey, R. G. and Carlaw, K. (1998a). *A Structuralist Assesment of Technology Policies: Taking Schumpeter Seriously on Policy*. Industry Canada. Ottawa. Working Paper Number 25.

Lipsey, R. G. and Carlaw, K. (1998b). Technology Policies in Neo-Classical and Structuralist-Evolutionary Models. *OECD Science Technology and Industry Review,* 22 (Special Issue on "New Rationale and Approaches in Technology and Innovation Policy")**,** 30-73.

Lipsey, R. G. and Carlaw, K. (2002). The Conceptual Basis of Technology Policy. *Department of Economics Discussion Papers.* Vancouver: Simon Fraser University

Lipsey, R. G., Carlaw, K. I. and Bekar, C. T. (2005). *Economic Transformations: General Purpose Technologies and Long Term Economic Growth,* New York, Oxford University Press.

Luukkonen, T. (2000). Additionality of EU framework programmes. *Research Policy,* 29**,** 711-724.

Malik, K., Georghiou, L. and Cameron, H. (2006). Behavioural Additionality of the UK Smart and Link Schemes. *In:* OECD (ed.) *Government R&D Funding and Company Behaviour: Measuring Behavioural Additionality.* Paris: OECD.

OECD (2006). Evaluation of Publicly Funded Research: Recent Trends and Perspectives. *In:* OECD (ed.) *OECD Science, Technology and Industry Outlook.* Paris: OECD Publishing.

Řezanková, H. (2009). Cluster analysis and categorical data. *Statistika***,** 216-232.

Steurs, G., Verbeek, A., Vermeulen, H. and Clarysse, B. (2006). A Look into the Black Box: What difference do IWT R&D grants make for their clients? *IWT-Studies***,** 93.

Part II Chapter 7

Evaluation in the Context of Structural Funds: Impact on Evaluation Culture and Institutional Build up

The aim of the specific case study is to examine if and in what ways the Structural Funds (SF) requirements and regulations related to evaluation influence the evaluation culture, institutional build up and good practice in evaluation in specific countries. It draws upon the results of the questionnaire survey carried out under the INNO-Appraisal study and the examination of the uptake of SF regulations in three countries, Greece, Malta and Poland with a case study approach.

The questionnaire survey results reveal that while SF requirements do lead to some specific characteristics in delivery and practice of evaluation, they do not tend to lead to higher quality evaluations. Additionally, even high quality SF evaluations do not necessarily have greater impact in terms of useful and dissemination of results.

The country cases examined provide possible explanations for these interesting findings. At the same time, they show that while SF regulations have caused positive impacts in terms of capacity and structure building, they still fall short in improving institutional learning and establishing sound evaluations systems in the countries examined.

Effie Amanatidou and Ioanna Garefi[81]

---

[81] The lead authors for this section of the report are as given above, with contributions from all other team members at different parts of its production, conclusions and interpretations have been approved and are shared by the group.

Table of Contents

Table of Contents 202

List of Tables 203

Table of Exhibits 203

Executive Summary 204

1 Introduction 206

2 Evaluation requirements according to the SF regulations and provisions 208

2.1 Requirements for Structures related to Evaluation 208

2.2 Requirements for evaluation design and execution 209

2.3 Requirements for certain types of evaluation 209

2.3.1 Ex-ante evaluation 210

2.3.2 Mid-term evaluation 211

2.3.3 Ex-post evaluation 212

2.4 Information and publicity 212

2.5 Selection of Evaluators and Quality Assurance 212

2.6 Changes in SF regulations 213

3 Delivery and good practice in SF vs. non-SF type evaluations 214

3.1 Do SF requirements lead to specific characteristics in delivery & practice of evaluation? 215

3.2 Do SF requirements lead to higher quality evaluations? 221

3.3 Do high quality SF evaluations have greater impact? 224

4 Implementation and Impacts of SF evaluation requirements in different contexts 226

4.1 Greece 226

4.1.1 Innovation System 226

4.1.2 Evaluation Culture 228

4.1.3 Conclusions 230

4.2 Malta 231

4.2.1 Innovation system 231

4.2.2 Evaluation culture 231

4.2.3 Conclusions 234

4.3 Poland 234

4.3.1 Innovation System 234

4.3.2 Evaluation culture 235

4.3.3 Conclusions 237

4.4 SF regulations: a need for institutional learning and structure building 238
4.5 Explaining the survey results through the case studies 239
5 Conclusions 239
References 241

List of Tables

Table 83: Quality standards for evaluation report and evaluation process 213
Table 84: Evaluation topics covered 216
Table 85: Evaluation topics per evaluation type 217
Table 86: Data analysis methods 218
Table 87: Data analysis methods per evaluation type 219
Table 88: Data collection methods 219
Table 89: Data collection methods per evaluation type 220
Table 90: Degree of compliance to quality characteristics 222
Table 91: Usefulness of recommendations 223
Table 92: Usefulness of Recommendations and use of external evaluators 223
Table 93: Quality characteristics and provision of evaluation in programme design (SF sample) 224
Table 94: Quality characteristics and provision of evaluation in programme design (non SF sample) 224
Table 95: Quality characteristic, usefulness of recommendations and dissemination (SF sample) 225
Table 96: Quality characteristic, usefulness of recommendations and dissemination (non SF sample) 226

Table of Exhibits

Exhibit 68: Logic flow-chart of the methodological approach 207

Executive Summary

The aim of the specific case study is to examine if and in what ways the Structural Funds (SF) requirements and regulations related to evaluation influence the evaluation culture, institutional build up and good practice in evaluation in specific countries.

It draws upon the results of the questionnaire survey carried out under the INNO-Appraisal study and also the examination of the uptake of SF regulations in three countries, Greece (a Southern European country) and two new Member States, Poland and Malta. **T**he specific countries are examined as indicative examples of how SF evaluation related regulations and provisions are implemented and affect evaluation practices in their specific contexts.

The case study collection and analysis of data, information and stakeholder views is guided by the following hypotheses:

- SF requirements may lead to specific characteristics in delivery and practice of evaluation
- SF requirements may lead to higher quality evaluations
- High quality SF evaluations may have greater impact
- SF regulations demand high standards on structures and processes that inevitable need some institutional learning and structure building

SF regulations do seem to lead to specific characteristics in the delivery and practice of evaluation. SF type evaluations tend to be built in the design phase of a programme/measure. They usually also meet the requirement to make the results publicly available through publication of the evaluation report. Recommendations mainly relate to the programme / measure being appraised in terms of design, management and implementation clearly reflecting the orientation of the SF evaluations.

SF requirements also seem to contribute to guiding the evaluation topics covered under the different evaluation types (ex-ante, interim, ex-post) as well as the data analysis methods used (but not the data collection methods). Yet, SF guidelines seem to more or less repeat what is suggested by international practice in evaluation and thus also followed by non SF type evaluations. This might be the reason why no major differences exist when studying the results within the same evaluation type (ex-ante, or ex-post for example) across the two populations (SF and non SF).

SF requirements do not seem to lead to higher quality appraisals and even high quality SF evaluations do not lead to high impact in terms of usefulness of recommendations and dissemination of results. However, the suggestion to use independent (external) evaluators does seem to contribute to higher quality SF evaluations.

The country cases provide possible explanations for the survey results. The fragmentation among the key actors in the national innovation system in Greece, for example, and the fact that there is only mechanistic abidance to SF regulations can explain why the results of SF evaluations are only to a limited degree discussed with government and wider stakeholders.

Abidance by the 'letter rather than the essence of the law' in combination with doubts about the suitability of the SF regulations to lead to high impact evaluations can explain the limited usefulness of recommendations as well as the fact that even high quality SF evaluations may not lead to high impacts in terms of usefulness and dissemination of results. The fact that SF regulations and quality standards are only suggested rather than imposed may explain why suggested quality criteria may not be applied in practice.

Finally the country cases show that while SF regulations have caused positive impacts in terms of capacity and structure building, they still fall short in improving institutional learning and establishing sound evaluations systems in the countries examined.

1 Introduction

Evaluations of programmes supported under Structural Funds are often compulsory and both the European Commission and the National authorities in charge of the funds issue guidelines about when and how to perform the evaluative exercises. This according to some is a mixed blessing. On the one hand it tends to routinise the decision to evaluate but it is doubtful whether it offers clear added value for programme managers beyond abiding by the SF requirements. On the other hand it can provide a much needed and welcome guidance both to the planning authorities and to the evaluation teams about the expected behaviours and results. This availability of guidance has been an important stimulus for the development of evaluation capacity. (Tavistock Institute, et. al. 2003)

Following the above statement the aim of the specific case study is to examine if and in what ways the Structural Funds (SF) requirements and regulations related to evaluation influence the evaluation culture, institutional build up and good practice in evaluation in specific countries.

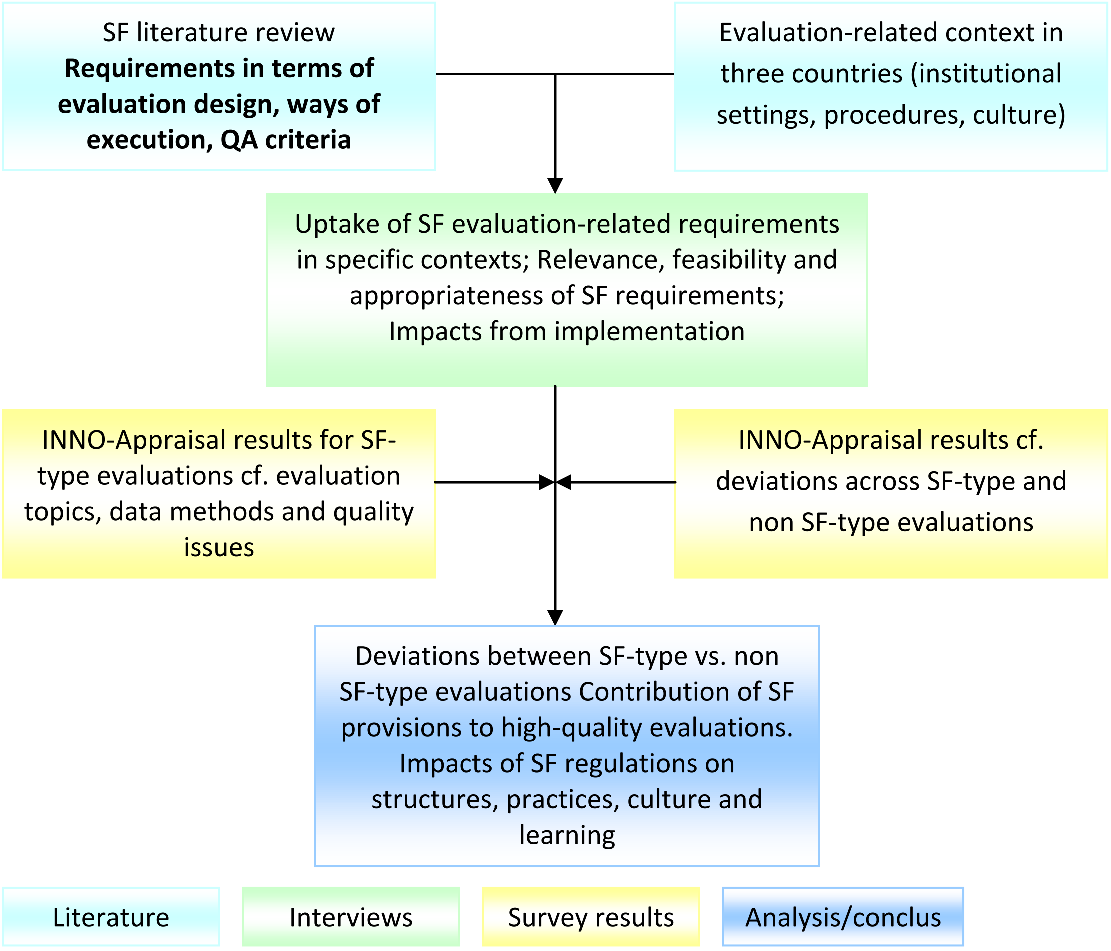
The case study covers three countries as indicative examples of how SF evaluation – related regulations and provisions are implemented and affect evaluation practices in their specific contexts: Greece (a Southern European country) and two new Member States, Poland and Malta. It has to be noted that the accession countries had no regulatory evaluation requirement to fulfil with respect to the European Structural Funds before 2006. In the meantime it was recommended to them that evaluation plans were prepared and budgets were allocated for evaluation activities during the period 2004-2006. (Tavistock Institute, et. al. 2003)

The case study collection and analysis of data, information and stakeholder views is guided by certain hypotheses that were used as indicative examples of what kind of analysis would be carried out:

- SF requirements may lead to specific characteristics in delivery and practice of evaluation
- SF requirements may lead to higher quality evaluations
- High quality SF evaluations may have greater impact
- SF regulations demand high standards on structures and processes that inevitable need some institutional learning and structure building

The methodology was based on desk research, telephone interviews and statistical elaboration of the data collected under the INNO-Appraisal survey. Below follows a logical flow-chart of the methodological approach.

Exhibit 68: Logic flow-chart of the methodological approach



SF literature review

Requirements in terms of evaluation design, ways of execution, QA criteria

Evaluation-related context in three countries (institutional settings, procedures, culture)

Uptake of SF evaluation-related requirements in specific contexts; Relevance, feasibility and appropriateness of SF requirements; Impacts from implementation

INNO-Appraisal results for SF-type evaluations cf. evaluation topics, data methods and quality issues

INNO-Appraisal results cf. deviations across SF-type and non SF-type evaluations

Deviations between SF-type vs. non SF-type evaluations Contribution of SF provisions to high-quality evaluations. Impacts of SF regulations on structures, practices, culture and learning

Literature

Interviews

Survey results

Analysis/conclus

Accordingly, the structure of the present report includes four sections. Section 2 summarises a literature review about the SF requirements and suggestions (in the form of guidelines, regulations or provisions) related to evaluation. These requirements are related to structures, evaluation design, the actual ways of execution covering specific evaluation types (i.e. ex-ante, mid-term / on going, and ex-post), as well as elements assuring high quality of evaluations.

The actual delivery of evaluation practices is then examined in Section 3 in two separate evaluation groups (SF type evaluations and non SF type evaluations) based on the INNO-Appraisal survey results. The latter group includes evaluations not bounded by SF regulations and carried out for single innovation measures rather than a set of measures. The aim is to examine the validity of three hypotheses:

1. Do SF requirements lead to specific characteristics in delivery & practice of evaluation?
2. Do SF requirements lead to higher quality evaluations?
3. Do high quality SF evaluations may have greater impact?

In turn, Section 4 examines the uptake of SF requirements and relevant suggestions in the context of three countries receiving SF support, i.e. Greece (a Southern European country) and two new Member States, Poland and Malta. Specific information was gathered through interviews based on an interview template (attached in Annex 7A) with experts in research and innovation programme / policy evaluation as well as SF regulations in each country.

The country cases aim to examine the fourth hypothesis, i.e. whether the demand of SF regulations in terms of high standards on structures and processes needs and/or leads to some institutional learning and structure building. The findings also provide possible explanations for some interesting survey results presented in section3.

Finally, Section 5 draws the main conclusions regarding the four research hypotheses as well as the general impacts of SF regulations on evaluation structures and institutional practices, culture and learning.

2 Evaluation requirements according to the SF regulations and provisions

Evaluation, along with monitoring and financial control, has enjoyed a prominent position in the regulations and provisions of Structural Funds (SF) over the years. Under the section titled 'Effectiveness' evaluation occupies a whole chapter involving general provisions, responsibilities of the Member States and the Commission as well as special sub-sections for ex-ante, mid-term and ex-post types of evaluations. (OJEC, 1999; OJEC, 2006)

Providing guidelines and suggestions on how to set up evaluation systems and conduct evaluations is not the only way SF assist the conduct of evaluations. The Community Strategic Guidelines for Cohesion 2007-2013, clearly state that the SF should support capacity building for public administrations at national, regional and local level, good policy design and implementation, including better lawmaking, evaluation and impact analysis of policy proposals, and regular screening of delivery mechanisms. The Guidelines put emphasis on the importance of appropriate, effective and transparent structures in central and regional administrations which are able to perform the tasks such as public procurement, financial control, monitoring, evaluation, and preventing and combating fraud and corruption. It is recognised that action should be supported in these directions. Thus, they encourage Member States to support good policy and programme design, monitoring, evaluation and impact assessment, through studies, statistics, expertise, and foresight, support for interdepartmental coordination and dialogue between relevant public and private bodies. (CEC, 2005)

The requirements of the SF regarding evaluation refer to several aspects such as structures, evaluation design and execution, the different types of evaluations (ex-ante, mid-term / on-going, and ex-post), the selection of evaluators and quality assurance criteria as well as ways to disseminate and publicize evaluation results.

2.1 Requirements for Structures related to Evaluation

SF provisions specify the structures to be created for the managing, monitoring and controlling SF interventions:

- A management authority responsible for the efficient, effective and correct management and implementation of an operational programme.
- A certification (previously paying) authority which draws up and sends to the Commission a certified inventory concerning expenditure and requests for payment. It must also certify the accuracy and the compliance of expenditure in terms of Community and national rules. It takes charge of accounting and assures the recovery of Community credits in the case of irregularities.
- An auditing (previously control) authority which is an operationally independent body designated by the Member State for each operational programme. It takes charge of the audits it carries out on the basis of an appropriate sample, writes up the annual control reports and offers an opinion on the audits carried out. The same authority can be assigned to a number of operational programmes.
- A follow-up committee, created for each operational programme by the Member State. It is presided over by a representative of the Member State or the management authority and is constituted according to a decision made by the Member State, and includes economic, social and regional partners. It assures the efficiency and the quality of the implementation of the operational programme.

In comparison with the previous programming period (2000-2006) the certification authority and the auditing authority replace the previous authority and control authority, while the responsibilities remain practically the same. (EU, 2007)

2.2 Requirements for evaluation design and execution

In preparing the documents to be submitted, MS should provide the following in relation with evaluation (OJEC, 1999):

- information on appropriations for preparing, monitoring and evaluation assistance (in the Community Support Framework Document);
- a description of the system for monitoring and evaluation and role of Monitoring Committee (in the Operation Programme document)
- the ex-ante evaluation itself, and monitoring indicators (in the Programme Complement);
- information on resources for preparation, monitoring, and evaluation assistance (in the Single Programming Document).

2.3 Requirements for certain types of evaluation

The Provisions for Structural Funds (OJEC, 1999; 2006) have a whole chapter dedicated to Evaluation. They dictate that Community structural assistance shall be the subject of ex-ante, mid-term and ex-post evaluation designed to appraise its impact with respect to the objectives set out and to analyse its effects on specific structural problems. They also allow for supplementary evaluations on the initiative of either the MS or the Commission after informing the MS.

Evaluations shall aim to improve the quality, effectiveness, and consistency of the assistance from the Funds and the, strategy and implementation of operational programmes with, respect to the specific structural problems affecting the, Member States and regions concerned, while taking account of the objective of sustainable development and of the relevant Community legislation concerning environmental impact and strategic environmental assessment. (OJEC, 2006)

Evaluations may be of a strategic nature in order to examine the evolution of a programme or group of programmes in relation to Community and national priorities, or of an operational nature in order to support the monitoring of an operational programme. (OJEC, 2006)

Adding to the above there are clear suggestions on how these evaluations should be carried out. Separate working documents[82] are provided with suggestions towards MS for several issues, for example on how to conduct ex-ante or mid-term evaluations or how to develop monitoring indicators. Reference to use existing guides on evaluation is made, especially the MEANS guide and its newest version EvalSED.

Below follows an indication of the specifications and suggestions made for each evaluation type.

2.3.1 Ex-ante evaluation

The ex-ante evaluation is an interactive process providing judgement and recommendations by experts, separately from the planners, on policy or programme issues. It falls under the responsibility of the MS. The objective is to improve and strengthen the final quality of the Plan or Programme under preparation. The SF Provisions (OJEC, 1999) clearly specify the contents that the ex-ante evaluation should have:

- an analysis of the strengths, weaknesses and potential of the Member State, region or sector concerned,
- assessment of the consistency of the strategy and targets selected with the specific features of the regions or areas concerned,
- the expected impact of the planned priorities for action, quantifying their specific targets in relation to the starting situation,
- an ex-ante evaluation of the socio-economic situation with particular emphasis in employment and human resources,
- an ex-ante evaluation of the environmental situation including a description, quantified as far as possible, of the existing environmental situation and an estimate of the expected impact of the strategy and assistance on the environmental situation,
- an ex-ante evaluation of the situation in terms of gender equality and an estimate of the related impact of the strategy and assistance,
- verification of the relevance of the proposed implementing and monitoring arrangements and consistency with Community policies.

---

[82] See for example EC, 2006a; 2006b; 2006c; 2007; 1999a; 1999b; CEC, 2000.

The ex-ante evaluation contributes to a better understanding of the following evaluation topics:

- The relevance of the existing strategy or the need for amendment.
- The effectiveness of existing policy delivery instruments.
- The critical factors affecting implementation and effectiveness.
- The types of problem in terms of policy evaluability and monitoring.

It is of major importance to ensure linkages and interactivity of ex-ante evaluation and policy development.

2.3.2 Mid-term evaluation

Mid-term evaluation, according to SF Provisions (OJEC, 1999) is to examine, in the light of the ex-ante evaluation, the initial results of the assistance, their relevance and the extent to which the targets have been attained. It shall also assess the use made of financial resources and the operation of monitoring and implementation. The key concerns are:

- Previous Evaluation Results;
- Continuing Validity of Analysis of Strengths, Weaknesses and Potential;
- Continuing Relevance and the Consistency of the Strategy;
- The Quantification of Objectives – Outputs, Results and Impacts;
- Effectiveness and Efficiency To Date and Expected Socio-Economic Impacts and, on this basis, Evaluation of the Policy and Financial Resources Allocation;
- Quality of Implementation and Monitoring Arrangements; and
- The Results for the Indicators agreed for the Performance Reserve.

Each mid term evaluation should be guided by a Steering Group representative of the monitoring committee for the form of assistance. The Steering Group's role is to develop the terms of reference for the evaluation, select the evaluators, guide the evaluation, give feedback on the first draft and approve it for quality on completion.

The evaluation should be organised to ensure that full use is made of the monitoring information which has been gathered over the two to three years of implementation and that the evaluators do not engage in unnecessary work in this regard.

The SF Provisions (OJEC, 1999) also dictate that specific criteria and indicators are set by the Commission in partnership with the MS to assess the performance of each OP or SPD. These indicators are based on an indicative list of indicators proposed by the Commission which are then quantified in the annual implementation and mid-term evaluation reports.

The importance of well conceived indicators and monitoring systems is strongly emphasised in relevant evaluation guidebooks. As The Guide (Tavistock Institute, et. al. 2003) notes "well conceived indicators and monitoring systems are a powerful adjunct to evaluation. Very often evaluators depend on monitoring systems which are indicator based. If these are not put in place early in the programme design cycle it may be too late to create such monitoring systems later on."

2.3.3 Ex-post evaluation

The Commission shall carry out an *ex post* evaluation for each objective in close cooperation with the Member State and managing authorities. *Ex post* evaluation shall cover all the operational programmes under each objective and examine the extent to which resources were used, the effectiveness and efficiency of Fund programming and the socio-economic impact. It shall draw conclusions for the policy on economic and social cohesion. It shall identify the factors contributing to the success or failure of the implementation of operational programmes and identify good practice. (OJEC, 2006)

While the mid-term evaluation falls under the responsibility of the managing authority in cooperation with the Commission and the MS, ex-post evaluation is the responsibility of the Commission in collaboration with the two other parties. Both of them are dictated to be executed by independent assessors.

2.4 Information and publicity

Provisions about information and publicity activities refer not only to SF activities but also evaluation results. The Member States should guarantee that information and the publicity of the Funds' activities concerning citizens and beneficiaries will be delivered. The objective is to highlight the role of the Union and to guarantee transparency. (EU, 2007) Specific note is also made that the evaluation results are made available to the public on request. Where possible, summaries should be placed on the internet. (OJEC, 1999) The Commission regards it as good practice to make public the entire evaluation report. (CEC, 2000) As the Guide (Tavistock Institute, et. al. 2003) notes '*Evaluation is wasted without communication of findings'.*

2.5 Selection of Evaluators and Quality Assurance

It is explicitly stated that evaluations shall be carried out by experts or bodies, internal or external, functionally independent of the local authorities responsible for managing SF by means of a competitive tendering process, either open or closed. (OJEC, 1999)

Furthermore, the Commission invites the competent authorities to ensure the quality of the evaluations. Authorities lacking quality standards are suggested to consult various guides like the MEANS publication which has been updated to EvalSED[83], or various other guidebooks like those published by DG Budget[84] and other Commission Directorates, other international organisation like OECD[85], or consultants[86]. Such guides provide useful advice and suggestions covering all the phases, types, and elements of an evaluation exercise.

---

[83] http://ec.europa.eu/regional_policy/sources/docgener/evaluation/evalsed/guide/index_en.htm.

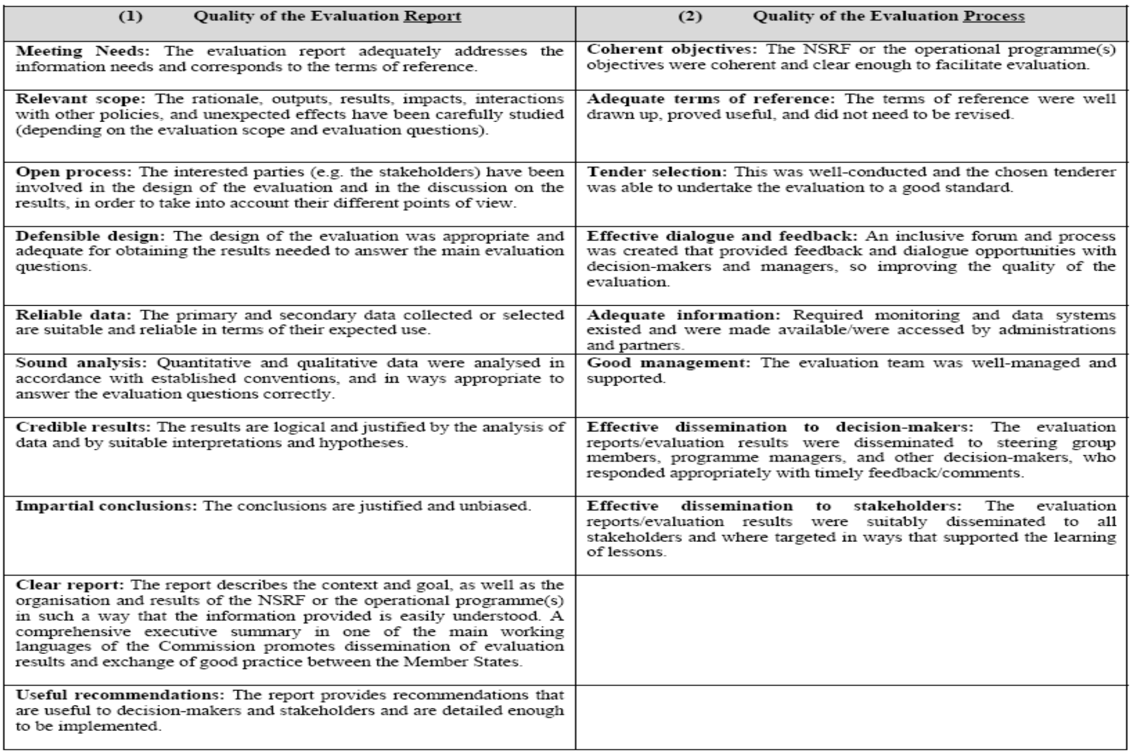
[84] http://ec.europa.eu/budget/library/publications/financial_pub/eval_activities_en.pdf.

[85] For example OECD (1999), 'Improving Evaluation Practices. Best Practice Guidelines for Evaluation and Background Paper.

[86] For example LL&A, PREST, ANRT and Reidev Ltd (2006), 'Supporting the monitoring and evaluation of innovation programmes. A study for DG Enterprise and Industry' Final Report, January 2006; SQW (2009), 'Pushing the boundaries of Impact Evaluation, Report on Knowledge Development Possibilities', April 2009.

Additionally, based on the above and other works certain quality standards are proposed for both the evaluation report and process under the SF provisions such as those presented in the Working Paper 5 for on-going evaluations.

Table 83: Quality standards for evaluation report and evaluation process



(1) Quality of the Evaluation Report	(2) Quality of the Evaluation Process
Meeting Needs: The evaluation report adequately addresses the information needs and corresponds to the terms of reference.	**Coherent objectives:** The NSRF or the operational programme(s) objectives were coherent and clear enough to facilitate evaluation.
Relevant scope: The rationale, outputs, results, impacts, interactions with other policies, and unexpected effects have been carefully studied (depending on the evaluation scope and evaluation questions).	**Adequate terms of reference:** The terms of reference were well drawn up, proved useful, and did not need to be revised.
Open process: The interested parties (e.g. the stakeholders) have been involved in the design of the evaluation and in the discussion on the results, in order to take into account their different points of view.	**Tender selection:** This was well-conducted and the chosen tenderer was able to undertake the evaluation to a good standard.
Defensible design: The design of the evaluation was appropriate and adequate for obtaining the results needed to answer the main evaluation questions.	**Effective dialogue and feedback:** An inclusive forum and process was created that provided feedback and dialogue opportunities with decision-makers and managers, so improving the quality of the evaluation.
Reliable data: The primary and secondary data collected or selected are suitable and reliable in terms of their expected use.	**Adequate information:** Required monitoring and data systems existed and were made available/were accessed by administrations and partners.
Sound analysis: Quantitative and qualitative data were analysed in accordance with established conventions, and in ways appropriate to answer the evaluation questions correctly.	**Good management:** The evaluation team was well-managed and supported.
Credible results: The results are logical and justified by the analysis of data and by suitable interpretations and hypotheses.	**Effective dissemination to decision-makers:** The evaluation reports/evaluation results were disseminated to steering group members, programme managers, and other decision-makers, who responded appropriately with timely feedback/comments.
Impartial conclusions: The conclusions are justified and unbiased.	**Effective dissemination to stakeholders:** The evaluation reports/evaluation results were suitably disseminated to all stakeholders and where targeted in ways that supported the learning of lessons.
Clear report: The report describes the context and goal, as well as the organisation and results of the NSRF or the operational programme(s) in such a way that the information provided is easily understood. A comprehensive executive summary in one of the main working languages of the Commission promotes dissemination of evaluation results and exchange of good practice between the Member States.	
Useful recommendations: The report provides recommendations that are useful to decision-makers and stakeholders and are detailed enough to be implemented.	

Source: EC, 2007, p. 18.

2.6 Changes in SF regulations

More responsibility with the MS

In comparison with the previous programming period (2000-2006) there is increased confidence to the Member States' control systems when they are the main financial contributors to the development programmes (Art. 74). If the trustworthiness of the projects is assured from the beginning of the period, audits of the Commission services will only be carried out in exceptional circumstances. (EU 2007) The new regulations (2007-2013) require that an Audit Authority is established for every Operational Programme separate from the Certifying Authority (replacing the Paying authority). More responsibility lies with the MS now than the Commission to preserve the audit trail and conduct the necessary checks and controls. (Smail, 2007)

Less obligatory evaluations

The new regulations offer greater flexibility by reducing the number of obligatory evaluations for the new programming period, 2007-2013. In the previous programming period (2000-2006) it was necessary that each intervention was the subject of ex-ante, mid-term and ex-post evaluation. Now ex-ante evaluations need to be carried out for each convergence objective programme while for each regional, competitiveness, employment and European territorial cooperation objective it is the

MS that can choose the level of evaluation according to their needs (programme, theme, funds). Mid-term evaluation can also be carried out according to needs. (EU, 2007)

From mid-term to on-going evaluations

Regulation 1083/2006 provides for a shift from a concept of mid-term evaluation driven by regulatory imperatives towards a more flexible, demand-driven approach to evaluation: on-going evaluation. On-going evaluation is a process taking the form of a series of evaluation exercises. Its main purpose is to follow on a continuous basis the implementation and delivery of an operational programme and changes in its external environment, in order to better understand and analyse outputs and results achieved and progress towards longer-term impacts, as well as to recommend, if necessary, remedial actions. The proposed approach emphasises the need for stronger links between monitoring and evaluation on the one hand, and on the other, between these two – very often - interlinked exercises and decision-making.

More information and publicity

In comparison with the previous period, 2000-2006, rules concerning information and publicity have been strengthened, notably as far as the follow-up to communication plans is concerned, as well as information concerning (potential) beneficiaries and the obligation of beneficiaries to communicate to the public the contribution the Funds have made to different projects. (EU, 2007)

3 Delivery and good practice in SF vs. non-SF type evaluations

Drawing upon the findings of the statistical elaboration of the data from the INNO-Appraisal survey this section answers the following questions:

- What is actually delivered in terms of evaluation practice in SF-type of evaluations and non SF-type?
- What do the INNO-Appraisal data say about the characteristics of high quality SF type evaluations? What makes SF evaluations successful in terms of usefulness, dissemination and uptake of results?) What do they say about non SF – type evaluations? What are the differences between in two groups (SF vs. non SF type evaluations) in terms of high quality characteristics?
- Do the results of the INNO-Appraisal survey regarding high quality SF type evaluations match what the SF requirements refer to?

The aim is to examine the validity of the first three of the hypotheses mentioned in the introduction:

- Do SF requirements lead to specific characteristics in delivery & practice of evaluation?
- Do SF requirements lead to higher quality evaluations?
- Do high quality SF evaluations may have greater impact?

The findings presented are based on the following cases:

- SF Sample A (INNO-Appraisal survey questions C.xxx): 38 cases

- SF Sample C (INNO-Appraisal survey questions D-F.xxx): 30 cases
- Non SF Sample A (INNO-Appraisal survey questions C.xxx): 133 cases
- Non SF Sample C (INNO-Appraisal survey questions D-F.xxx): 102 cases

3.1 Do SF requirements lead to specific characteristics in delivery & practice of evaluation?

The SF regulations suggest the **use of independent assessors** (either internal or external bodies) through the use of competitive tendering processes (either open or closed). The survey results showed that while the use of external bodies is preferred to by 51%, 32% of the respondents stated they used internal bodies while 16% said they applied a mixed approach.

Despite the fact that SF regulations do not point to a clear preference towards external bodies to ensure independence in a clear cut way, one might expect such a clear preference in the results. However, this was not the case. In fact the opposite was shown in the sense that a clear preference towards external bodies appeared for the majority (84.1%) of the non-SF type evaluations.

Referring to the **tendering process** there is a clear preference towards open procedures for 40% of the SF-type evaluations but there was also a small number of cases where no tendering process was applied (8%)

The SF-type evaluations are split among **ex-ante** (31.6%), **ex-post** (18.4%) and **interim** (31.6%) type evaluations. In the non SF type evaluations the dominant type is interim evaluations (46.2%) followed by ex-post (31.1%) and accompanying (13.6%). This should be taken into account in explaining any differences in the evaluation topics covered or data selection and analysis methods as they might be due to the dominance of interim evaluations in the non SF type group or of ex-ante evaluations in the SF group rather than in genuine differences in characteristics of SF and non SF type evaluations.

Slightly more than half (51.4%) of the SF type evaluations are considered **formative**, which can be explained by the majority of the ex-ante in the respective group. Around 20% of them are considered as both summative and formative. In the non SF type cohort more evaluations are considered as both summative and formative (36.4%) – which can be attributed to the dominance of interim evaluations in the respective group - while 38.6% are considered as formative, and 22% as summative.

As expected in the SF cases the appraisal was a **condition** of internal **(co)sponsorship** for the majority of the cases (77.8%) with most of them reporting the EU or Structural Funds as the sponsor. In the non SF type this was the case for only 9.6%.

SF regulations require the planning of the evaluation of the interventions even at the design phase of the interventions foreseen. Thus, in the great majority (84.2%) of the SF cases the appraisals were **planned during the design phase** of the measure. This was the case for less (62.4%) of the non SF cases.

Referring to the **evaluation topics covered**, there appear to be less striking differences than one would expect. In fact the issue of 'internal / external consistency', 'goal attainment/effectiveness', ‘outputs, outcomes and impacts’, ‘programme implementation efficiency’ and ‘policy / strategy development’ are topics covered by significant shares in both cohorts. Naturally, ‘internal consistency’ is slightly more chosen by the SF group given that they form specific elements to address in SF type evaluations. On the other hand, ‘behavioural additionality’ is much more addressed in non SF type evaluations while gender and minority issues are analysed more in SF type given also the fact that they form explicit sections in the relevant documents that have to be filled in and submitted.

Table 84: Evaluation topics covered

Evaluation Topics	SF type (%)	Non SF type (%)
External Consistency	86,10%	76,40%
Internal Consistency	91,40%*	77,50%*
Coherence/Complementarity	67,60%	60,80%
Goal Attainment/Effectiveness	81,30%	91,30%
Outputs, Outcomes and Impacts	86,80%	92,90%
Quality of Outputs	60,00%	56,70%
Value for Money/Return on Investment/Cost-Benefit	38,50%	24,80%
Programme Implementation Efficiency	79,30%	75,40%
Project Implementation Efficiency	54,80%	45,00%
Input Additionality	56,30%	48,30%
Output Additionality	55,90%	48,30%
Behavioural Additionality *	33,30%*	54,70%*
Policy/Strategy Development	80,00%	75,20%
Gender issues *	55,90%*	14,50%*
Minority issues *	19,40%*	3,50%*

(*) Differences across SF and non SF samples significant at 95% significance level

It might be expected that there are clearer differences in the evaluation topics among **ex-ante, mid-term and ex-post** SF type evaluations given the particular instructions on how to conduct these types of evaluations under the SF framework.

Indeed, differences become clearer when looking at the evaluation topics per evaluation type under the SF sample. For example, interim evaluations tend to address more issues of effectiveness, internal consistency, outputs and impacts and programme implementation efficiency while the ex-ante type mainly focuses on the aspects of consistency, and complementarity while also addressing slightly more policy / strategy development. Ex-ante evaluations also take more into account gender and minority issues as they form specific sections in the relevant SF documents that have to be prepared and submitted. Ex-post SF evaluations address more issues of effectiveness and outputs while accompanying SF evaluations cover a variety of issues combining both an interim and ex-post orientation.

Within the non SF sample the topics covered present less differences across the evaluation types rather than in the SF sample. This may be attributed partly to the clear guidelines provided under the SF framework about the focus that each of the evaluation types should have.

For each evaluation type across the two samples however (e.g. ex-ante SF vs. ex-ante non SF) the picture is rather uniform. For instance the issues addressed by interim evaluations are more or less the same. In fact no statistical significant differences exist in the respective shares across the two samples (SF vs. non-SF) either for the interim or the accompanying evaluations. In the case of ex-ante evaluations while the same topics are addressed across the two populations it is in the SF sample that they are used by significantly larger shares.

The ex-post evaluations also present commonalities across the two populations such as a clear preference in effectiveness and impacts but there are differences as well. The ex-post non SF type evaluations address much more than their SF equivalents issues of consistency and coherence / complementarity as well as issues of value for money, programme efficiency and input additionality).

Table 85: Evaluation topics per evaluation type

	'SF' sample				**'non SF' sample**			
Evaluation topics	Ex ante	Acc	Interim	Ex post	Ex ante	Acc	Interim	Ex post
External Consistency	33,33%*	16,67%	27,78%	5,56%*	4,72%*	11,02%	36,22%	22,83%*
Internal Consistency	31,43%*	17,14%	34,29%	5,71%*	7,03%*	11,72%	36,72%	20,31%*
Coherence/Complementarity	32,35%*	8,82%	17,65%	5,88%*	5,04%*	9,24%	27,73%	18,49%*
Goal Attainment/Effectiveness	6,25%	18,75%	37,50%	15,63%	4,80%	13,60%	44,00%	27,20%
Outputs, Outcomes and Impacts	18,42%	15,79%	31,58%	18,42%	6,35%	14,29%	42,86%	27,78%
Quality of Outputs	6,67%	20,00%	13,33%	16,67%	2,52%	9,24%	21,85%	21,85%
Value for Money/RoI/Cost-Benefit	7,69%	7,69%	15,38%	3,85%*	0,00%	2,59%	8,62%	13,79%*
Programme Implem. Efficiency	20,69%*	17,24%	31,03%	6,90%*	4,80%*	9,60%	35,20%	24,00%*
Project Implementation Efficiency	3,23%	16,13%	19,35%	12,90%	1,68%	5,04%	20,17%	16,81%
Input Additionality	9,38%	15,63%	18,75%	9,38%*	2,59%	5,17%	17,24%	22,41%*
Output Additionality	8,82%	17,65%	11,76%	14,71%	2,61%	6,96%	18,26%	20,00%
Behavioural Additionality	0,00%	10,00%	13,33%	6,67%	3,42%	8,55%	21,37%	21,37%
Policy/Strategy Development	25,71%*	14,29%	25,71%	11,43%	7,03%*	11,72%	36,72%	18,75%
Gender issues	26,47%*	14,71%	11,76%	0,00%	2,59%*	3,45%	6,90%	0,86%
Minority issues	19,35%*	0,00%	0,00%	0,00%	0,89%*	1,79%	0,00%	0,89%

(*) Differences in same sub-group across SF and non SF samples significant at 95% significance level

The two different groups of evaluations (SF and non SF) do not present striking differences in terms of the **data analysis** methods used. In general 'descriptive statistics' and 'context analysis' stand out in both SF and non SF type evaluations. 'Document analysis' comes next in either case. There are some differences in relation to the next in rank methods. The SF sample is characterised by an increased use of input/output analysis, before/after group comparison and cost/benefit approaches, while the non SF sample applies much more case study analysis as expected in mainly interim and ex-post evaluations.

Table 86: Data analysis methods

Data analysis	SF type (%)	Non SF type
Case Study Analysis	22,90%*	46,50%*
Network Analysis	14,80%	17,90%
Econometric Analysis	31,40%	20,30%
Descriptive Statistics (e.g. uptake analysis)	69,40%	77,40%
Input/Output Analysis	51,40%*	18,70%*
Document Analysis	61,30%	49,20%
Context Analysis	86,10%*	61,50%*
Before/After Group Comparison Approach	24,20%*	6,60%*
Control Group Approach	19,20%	20,20%
Counter-Factual Approach	12,50%	23,80%
Cost/Benefit Approach	50,00%*	15,40%*

(*) Differences across SF and non SF samples significant at 95% significance level

One might also expect clear differences in the data analysis methods used in the different types of evaluations (accompanying, ex-ante, ex-post and interim) in either sample. Nevertheless, the survey findings show that the differences are not that marked within either the SF or the non SF sample. Within the SF sample it is ‘context analysis’, ‘descriptive statistics’, ‘input/output analysis’ and cost/benefit analysis’ that dominate in 3 out of the 4 evaluation types. Within the non SF group the same picture emerges with ‘context analysis’, and descriptive featuring in high share position across all four evaluation types. However, as noted above ‘case study analysis’ also features as significant in the three out of the four evaluation types.

When the findings per evaluation type are examined across the two samples the results are not that different either. For example, the data analysis methods mainly used in ex-ante evaluations across both samples (SF and non SF) are ‘document’ and ‘context analysis’ and ‘descriptive statistics’, although they are used by significant larger shares in the SF sample. The interim type evaluations also present similarities across both samples with ‘case studies’, ‘descriptive statistics’, ‘document’ and ‘context analysis’ featuring as top choices. Ex-post evaluations use mainly ‘descriptive statistics’ and ‘context analysis’ in both cases but the non SF group prefers more ‘document analysis’ as well as ‘case study analysis’ and ‘counter-factual approach’ Accompanying evaluations use ‘context analysis’ across both groups while ‘econometric analysis’ and ‘cost/benefit approach’ is used significantly more in the SF group.

The slightly different picture that emerges across the different evaluation types within the SF sample suggests that the SF requirements do contribute to guiding the data analysis methods. The uniformity, however, under each evaluation type (ex-ante, interim, ex-post) across the two samples (SF, non SF) suggests that the SF requirements repeat more or less what is advisable by international practice and thus followed by the non SF sample as well. Thus, while SF requirements do guide the data analysis in the different evaluation types, this guidance does not differ from what is usually followed in non SF type evaluations.

Table 87: Data analysis methods per evaluation type

	'SF' sample				'non SF' sample			
Data analysis methods	**Ex-ante**	**Acc**	**Interim**	**Ex post**	**Ex-ante**	**Acc**	**Interim**	**Ex post**
Case Study Analysis	0,00%	5,71%	14,29%	2,86%*	0,78%	7,81%	24,22%	13,28%*
Network Analysis	0,00%	7,41%	0,00%	3,70%	0,00%	2,46%	6,56%	9,02%
Econometric Analysis	2,86%	17,14%*	5,71%	2,86%	0,82%	1,64%*	8,20%	9,84%
Descriptive Statistics	13,89%*	8,33%	33,33%	11,11%	2,44%*	13,82%	38,21%	21,95%
Input/Output Analysis	5,71%	14,29%	14,29%	14,29%	1,64%	4,10%	6,56%	6,56%
Document Analysis	29,03%*	3,23%	22,58%	3,23%*	5,51%*	6,30%	22,05%	14,96%*
Context Analysis	33,33%*	16,67%	19,44%	13,89%	4,13%*	10,74%	25,62%	19,83%
Before/After Group Compar.	9,09%	9,09%	3,03%	0,00%*	0,83%	2,48%	0,00%	3,31%*
Control Group Approach	7,69%	0,00%	3,85%	7,69%	0,00%	1,63%	9,76%	8,94%
Counter-Factual Appr.	0,00%	0,00%	8,33%	4,17%*	0,00%	2,48%	7,44%	14,05%*
Cost/Benefit Approach	17,65%*	14,71%*	2,94%	11,76%	0,82%*	1,64%*	5,74%	7,38%

(*) Differences in same sub-group across SF and non SF samples significant at 95% significance level

Data collection methods present more differences across the SF and non SF samples. 'Existing surveys/databases' comes first in the SF group with 'document search' and 'monitoring data' following. The non SF results give priority to 'interviews', 'monitoring data' and then 'participant surveys'. This is understandable as the ex-ante and interim evaluations dominating the SF group are usually based more on secondary data while the interim and ex-post evaluations dominating the non SF group are primarily based on primary data collection.

Table 88: Data collection methods

Data collection	SF type (%)	Non SF type
Existing Surveys/Databases	92,10%*	63,00%*
Participant Surveys	47,20%*	69,80%*
Non-participant Surveys	16,00%	26,70%
Interviews	38,90%*	86,40%*
Focus Groups/Workshops/Meetings	50,00%	49,60%
Peer Reviews	12,50%	19,80%
Technometrics / Bibliometrics Search	0,00%	2,50%
Document Search	82,40%*	59,20%*
Monitoring Data	82,40%	78,40%

(*) Differences across SF and non SF samples significant at 95% significance level

The situation is similar and even more uniform among the different evaluation types. Within the SF group 'monitoring data' and 'existing surveys' feature in all evaluations types, while 'focus groups' and 'document search' feature in 3 out of 4 evaluation types. The differences refer mainly to the increased use of 'interviews', 'focus groups and meetings' in the ex-ante evaluations. This is expected as these are the usual means to discuss current situation and decide upon appropriate design of priorities and measures.

The other point of difference refers to the increased use of 'participant surveys' in interim evaluations which is again expected as this is the usual methods applied in evaluating the progress, as well as the outcomes and impacts of interventions. The use of 'participant surveys' along with 'interviews and 'focus groups and meetings' would be expected to be increased also in the case of ex-post evaluations. The fact that this is not the case might be explained from the few cases of ex-post SF evaluations. Other than that it is the ‘existing surveys’, ‘document search’ and ‘monitoring data’ that are mostly used in all evaluation types in the SF group.

The situation is not at all different in the non SF group where ‘monitoring data’, ‘existing surveys’ but also ‘interviews’ feature across all types of evaluations, while ‘document search’, ‘focus groups’ and ‘participants surveys’ in 3 out of 4 evaluation types.

Examining the findings under each evaluation type across the two groups, the main difference is the clear preference for ‘participant surveys’, ‘interviews’ and ‘focus groups, workshops and meetings’ for the non SF ex-post type evaluations.

The uniformity that the evidence reveals across the different evaluation types in the SF group suggests that the SF regulations / requirements do not make a real difference in terms of data collection methods. This is also reinforced by the fact that no major differences appear between the SF and non SF samples.

Table 89: Data collection methods per evaluation type

	‘SF’ sample				**‘non SF’ sample**			
Data collection methods	Ex-ante	Acc	Interim	Ex post	Ex-ante	Acc	Interim	Ex post
Existing Surveys/Databases	28,95%*	15,79%	26,32%	18,42%	3,17%*	8,73%	28,57%	21,43%
Participant Surveys	5,56%	8,33%	30,56%	2,78%*	1,60%	10,40%	32,00%	25,60%*
Non-participant Surveys	8,00%	0,00%	4,00%	4,00%	1,68%	1,68%	14,29%*	8,40%
Interviews	13,89%	2,78%*	19,44%*	2,78%*	5,34%	12,98%*	42,75%*	23,66%*
Focus Groups/WS/Meetings	18,75%*	6,25%	18,75%	3,13%*	4,76%*	7,94%	20,63%	15,87%*
Peer Reviews	4,17%	0,00%	8,33%	0,00%*	0,00%	1,67%	10,00%	6,67%*
Technometrics / Bibliometrics	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,83%	1,67%
Document Search	29,41%*	2,94%	32,35%	14,71%	5,65%*	7,26%	27,42%	17,74%
Monitoring Data	20,59%*	14,71%	32,35%	11,76%	2,61%*	14,78%	38,26%	21,74%

(*) Differences in same sub-group across SF and non SF samples significant at 95% significance level

Impact types addressed present differences that may be attributed to the nature of SF interventions and consequent evaluations that are quite broad. Indeed, most of the impact types are addressed by significant shares of SF type evaluations with economic impacts first followed by technological and then by social impacts. In most cases impacts affected participants as well as the wider environment. The impact types addressed by the non SF cases refer to more focused evaluations studying primarily economic impacts and to a lesser degree technological impacts.

The **audiences** of the evaluations in the case of SF type are primarily programme managers and (co)sponsors as well as policy analysts and government officials. These audiences were selected in the vast majority of cases (>90%). In the case of non SF the two first groups were mostly chosen, i.e. programme managers and government officials with 100% and 97,6% respectively.

The suggestion made explicitly in SF provisions that evaluation results should be made **available to the public** was reflected in the result that the total number of SF type evaluations are made available through the web (89.5%) or hard copies (5.3%). This was the case for 87% of the non SF type cases with the web again being the most preferred means of publication (76.3%). Thus, the hypothesis that ‘broad publicity of evaluation results counts more as a quality / good practice element in SF rather than in non SF type evaluations’ seems to be correct to a small degree.

Recommendations in the SF sample mainly related to the programme / measure being appraised in terms of design, management and implementation clearly reflecting the orientation of the SF evaluations. In the non SF sample preferences covered more or less all available options with the one related the future programme / measures coming first, followed by the management and implementation of programme/measure appraised and by the changes to broader policy formulation and implementation.

Referring to the **consequences** of the appraisal again the differences are negligible. The majority of cases in either sample did not present the consequences listed in the questionnaire. ‘Minor re-design of the measure’ was the consequence in most of the SF and non SF type evaluations that did state a particular consequence, while the non SF group also selected the ‘expansion / prolongation of the measure’.

Concluding SF regulations do seem to lead to specific characteristics in the delivery and practice of evaluation. Abiding by the respective requirements SF evaluations tend to be built in the design phase of a programme/measure while also following the request to make the results publicly available. Recommendations mainly relate to the programme / measure being appraised in terms of design, management and implementation clearly reflecting the orientation of the SF evaluations.

The difference in the evaluation topics covered by the different evaluation types can be attributed to the clear guidelines about the focus that each evaluation type should have. The slight differences across the different evaluation types within the SF sample also suggest that the SF requirements contribute to guiding the data analysis methods used (but not the data collection methods). Yet, since there are no major differences when looking at the different evaluation types across the two samples (SF and non SF) it is concluded that SF guidelines more or less repeat what is suggested by international standards in evaluation and thus also followed by non SF type evaluations.

3.2 Do SF requirements lead to higher quality evaluations?

The Inno-Appraisal survey addressed several elements of the quality of the appraisals. All of them can relate to quality standards proposed under the SF provisions[87]. For example, the degree to which the terms of reference are addressed refers to the quality standard titled ‘meeting needs’. Whether the design of the evaluation was appropriate is referred to as ‘defensible design’. The way the methods were applied and data analysed are referred to under ‘sound analysis’. The coverage of the wider context is mentioned under ‘clear report’ while the usefulness of recommendations is explicitly mentioned as a quality standard. The degree to which conclusions are based on the analysis refers to ‘credible results’ and ‘impartial conclusions’. Suitability of methods chosen can be implied under ‘sound analysis’ while the adequate documentation of information sources can be implied by the criterion ‘reliable data’. The degree to which results were discussed within government circles or with participants and broader stakeholders are explicitly mentioned as quality standards referring to the evaluation process.

---

[87] As mentioned for example in EC (2007), ‘The New Programming Period 2007-2013. Indicative Guidelines On Evaluation Methods: Evaluation During The Programming Period. Working Document No.5, DG Regional Policy, April 2007.

One might expect that the suggestion of **specific quality standards** for the SF type evaluations would lead to a greater degree of compliance of this type of evaluations in comparison with non SF type ones. However, the results of the Inno-Appraisal survey do not support this argument. On the contrary, it is the non SF type evaluations that present generally a higher score of compliance with most of the quality characteristics addressed than the SF type ones. Nevertheless, the relatively high scores in the SF sample should also be noted in several quality characteristics except the ones related to discussions of results and usefulness of recommendations.

Table 90: Degree of compliance to quality characteristics

Quality characteristic	Degree of compliance (medians) (1: Not at all, 5: Yes, definitely)	
	SF type	Non SF type
Suitability of methods chosen	4,00	5,00
Well documented information sources	4,00*	4,00*
Address of ToR	4,00	5,00
Analysis based on data	4,00*	4,00*
Design appropriate to objectives	4,00*	4,00*
Conclusions based on analysis	4,00	5,00
Satisfactory application of qualitative methods	4,00*	5,00*
Coverage of broader context	3,50	3,00
Satisfactory application of quantitative methods	4,00*	5,00*
Discussion within government circles	2,00*	4,00*
Discussion with participants / stakeholders	2,00*	4,00*
Usefulness of recommendations (^)	2,4^	3,1^

(^) Average of the medians of the five available options.

(*) Differences across SF and non SF samples significant at 95% significance level. Estimation of significance of differences is based on the means and not the medians. This is why significant differences may appear even in cases where the median is the same across the two samples (SF and non SF).

The most striking differences across the two samples appear in the characteristics related to the application of qualitative and quantitative methods, and the degree to which results are discussed with government of stakeholders. The latter is particularly interesting as dissemination of results to decision-makers and stakeholders is explicitly mentioned as a quality standard of the evaluation process in SF regulations.

This also happens in the case of **usefulness of recommendations**. This quality criterion scores particularly low in terms of compliance of SF type evaluations although it is also an explicit quality criterion in SF evaluation guides. Overall, these two elements (discussion of results and usefulness of recommendations) seem to be the two most ‘overlooked’ quality standards in SF type evaluations.

The usefulness of the recommendations was not considered particularly high in either sample, SF or non SF. The SF sample indicated that the relatively most useful recommendations were related to future programmes / measures although most of the recommendations were oriented towards the programme being appraised. The non SF group noted the changes in relation to policy formulation as significantly more useful than the SF group. Nevertheless, as already indicated above, non SF sample recommendations were considered more useful than recommendations in the SF sample.

Table 91: Usefulness of recommendations

Recommendations	Usefulness (median)	
	SF	Non SF
Changes to the design of the programme/measure appraised	3,00	3,00
Changes to the management and implementation of the programme/measure appraised	3,00*	3,50*
Changed to the design, management and implementation of future programmes/measures	4,00	4,00
Changes to the design, management and implementation of contemporaneous programmes/measures	1,00*	2,00*
Changes to broader policy formulation and implementation	1,00*	3,00*

(*) Differences across SF and non SF samples significant at 95% significance level. Estimation of significance of differences is based on the means and not the medians. This is why significant differences may appear even in cases where the median is the same across the two samples (SF and non SF).

The suggestion to use **independent (external) evaluators** might also contribute to higher quality SF evaluations. Indeed evidence seems to suggest that this may be true. All the recommendations related to the programme being appraised or future programmes were more useful when coming from evaluations assigned to external bodies. Thus, the choice of external bodies does make a difference in terms of usefulness of recommendations in the SF sample.

However, abiding by the general SF suggestion of using independent evaluators does not lead to increased usefulness of recommendations when compared with the non SF sample. In all types of recommendations, usefulness scores are at similar levels when evaluations are carried out by external bodies in the SF and the non SF samples.

Table 92: Usefulness of Recommendations and use of external evaluators

Recommendations	Usefulness (Median)			
	SF sample		Non SF sample	
	External bodies	Any body	External Body	Any body
Changes to the design of the programme/measure appraised	3,00	1,00	3,00	5,00
Changes to the management and implementation of the programme/measure appraised	3,00	2,00	3,00	5,00
Changed to the design, management and implementation of future programmes/measures	4,00	2,00	4,00	3,00
Changes to the design, management and implementation of contemporaneous programmes/measures	1,00	1,00	2,00	2,00
Changes to broader policy formulation and implementation	1,00	1,50	3,00	2,00

The great majority of the SF evaluations were planned during the design phase of the measure. Thus, it is purposeful to examine whether e**valuations that are built in** the design of a programme tend to be of better quality.

There is no indication that this is true for either the SF or the non SF population. Evaluations that are foreseen in the design of the programme, accounting for the vast majority (84.2%) of the SF cases, do not present different scores (on a 1-5 scale) on the various quality characteristics from the average SF case. Nor do they present significant differences from the SF evaluations that are not in-built.

Table 93: Quality characteristics and provision of evaluation in programme design (SF sample)

Quality characteristic	Median (appraisals not in-built: 3.3% of total)	Median (appraisals in-built: 96.7% of total)	Median (overall SF)
Address of ToR	4,00	4,00	4,00
Appropriate design	4,00	4,00	4,00
Suitability of methods chosen	4,00	4,00	4,00
Application of qualitative methods	3,00	4,00	4,00
Application of quantitative methods	3,00	4,00	4,00
Documentation of info sources	3,00	4,00	4,00
Analysis based on data	4,00	4,00	4,00
Coverage of broader context	5,00	3,00	3,50
Conclusions based on analysis	5,00	4,00	4,00

Note: no significance tests can be calculated due to lack of adequate data in not-built appraisals

The same stands for the non SF population. The only significant difference appears in the application of qualitative methods, which scores higher in the case of in-built evaluations (presenting a means of 3.92 vs. 3.47 in the case of not in-built evaluations).

Table 94: Quality characteristics and provision of evaluation in programme design (non SF sample)

Quality characteristic	Median (appraisals not in-built: 25.5% of total)	Median (appraisals in-built: 74.5% of total)	Median (overall non SF)
Address of ToR	5,00	4,50	5,00
Appropriate design	4,00	4,00	4,00
Suitability of methods chosen	4,00	5,00	5,00
Application of qualitative methods	4,00*	4,00*	4,00
Application of quantitative methods	5,00	4,00	4,00
Documentation of info sources	4,00	5,00	5,00
Analysis based on data	4,00	5,00	5,00
Coverage of broader context	4,00	3,00	3,50
Conclusions based on analysis	5,00	5,00	5,00

(*) Differences across SF and non SF samples significant at 95% significance level. Estimation of significance of differences is based on the means and not the medians. This is why significant differences may appear even in cases where the median is the same across the two samples (SF and non SF).

Concluding, SF requirements do not seem to lead to high quality appraisals. While compliance to quality standards is relatively high it is not higher than the non SF sample. Two specific quality criteria seem to be particularly overlooked, namely the discussion of results and usefulness of recommendations. Nevertheless, the suggestion to use independent (external) evaluators does seem to contribute to higher quality SF evaluations.

3.3 Do high quality SF evaluations have greater impact?

One might expect that appraisals with high scores in quality characteristics would generally result in high impact in terms of dissemination and usefulness of recommendations. However, evidence does not seem to clearly support this argument.

In the SF sample it seems that high scores in quality characteristics lead only to moderate or lower usefulness of recommendations and degree of dissemination of results. The only exception is the recommendations related to future programmes where high quality evaluations seem to lead to higher than moderate usefulness of this type of recommendation. The opposite stands for the recommendations with regards to contemporaneous programmes or policy formulation.

Table 95: Quality characteristic, usefulness of recommendations and dissemination (SF sample)

Sample: SF	Usefulness of Recommendations (Median of 1-5 scale)						Dissemination	
(N='4s' or '5s' in quality)	Pgm design	Pgm implem.	Future pgm	Contemp. Pgm	Policy form.		Discuss gov't	Discuss stkhold.
Address ToR	3,00	3,00	4,00	1,00	2,00*		3,00*	3,00
Eval. Design	3,00	3,00	4,00	1,00	1,50*		2,00*	3,00*
Methods	3,00	3,00	4,00	1,00	2,00*		3,00*	3,00*
Qualitative	3,00	3,00	3,50	1,00*	1,50*		2,50	2,00*
Quantitative	3,00	3,00	4,00	-	2,50		2,50	2,00
Information	3,00	3,00*	4,00	1,00*	1,00*		2,00*	2,00*
Analysis	3,00	2,50	3,50	1,00*	1,00*		2,00*	2,00*
Context	2,00*	2,50	2,00*	1,00*	1,50*		2,00*	2,50*
Conclusions	3,00	2,00	4,00	1,00*	1,50*		2,00*	2,50*

(*) Differences across SF and non SF samples significant at 95% significance level. Estimation of significance of differences is based on the means and not the medians. This is why significant differences may appear even in cases where the median is the same across the two samples (SF and non SF).

In the case of non SF where the number of responses is significantly higher, the picture is more positive. Generally, high quality scores in all quality characteristics do seam to lead to higher than moderate degree of usefulness (≥ 3) in all types of recommendations (except in the case of recommendations related to contemporaneous programmes) and clearly high degree of dissemination or results.

In terms of individual quality characteristics that stand out in terms of increasing usefulness of recommendations and dissemination of results, the evaluation design and the degree to which the wider context is covered stand out. The latter result is interesting as compliance to the specific criterion (degree of coverage of the wider context) was only moderate in the non SF sample as shown above (Table 90: Degree of compliance to quality characteristics).

Table 96: Quality characteristic, usefulness of recommendations and dissemination (non SF sample)

Sample: Non SF	Usefulness of Recommendations (Median of 1-5 scale)						Dissemination	
(N='4s' or '5s' in quality)	Pgm design	Pgm implem.	Future pgm	Contemp. Pgm	Policy form.		Discuss gov't	Discuss stkhold.
Address ToR	3,00	3,00	4,00	2,00	3,00*		4,00*	4,00
Eval. Design	4,00	4,00	4,00	3,00	3,00*		4,00*	4,00*
Methods	3,00	3,00	4,00	3,00	3,00*		4,00*	4,00*
Qualitative	3,00	3,00	4,00	2,50*	3,00*		4,00	4,00*
Quantitative	3,50	3,50	4,00	2,50	3,00		4,00	4,00
Information	3,00	3,00*	4,00	2,00*	3,00*		4,00*	4,00*
Analysis	3,00	3,00	4,00	2,00*	3,00*		4,00*	4,00*
Context	4,00*	4,00	4,00*	3,00*	3,00*		4,00*	4,00*
Conclusions	3,00	3,00	4,00	2,00*	3,00*		4,00*	4,00*

(*) Differences across SF and non SF samples significant at 95% significance level. Estimation of significance of differences is based on the means and not the medians. This is why significant differences may appear even in cases where the median is the same across the two samples (SF and non SF).

Across the two samples the picture that stands out is that almost all quality characteristics improve significantly the usefulness of recommendations related to policy formulation as well as the degree results are discussed with both audiences examined (government and wider stakeholders) in the case of non SF type evaluations.

Concluding, the evidence does not strongly suggest that high quality SF evaluations lead to high impact in terms of usefulness of recommendations and dissemination of results.

4 Implementation and Impacts of SF evaluation requirements in different contexts

The present section answers the following questions:

1. Are the SF requirements relevant, feasible and appropriate to the contexts of the countries studied?
2. What are the impacts from implementing the SF requirements in the broader innovation system, on the quality of the evaluation and policy making systems in the countries examined? How much have the structures and institutional practices changed over time and how much is this consequence of the SF requirements? What do we see in terms of policy and evaluation learning?

It is mainly based on desk research and telephone interviews with experts in the three countries covered.

4.1 Greece

4.1.1 Innovation System

The Greek national innovation system was declared as being the weakest among the different EU countries and the OECD area. It has also been stated that up until now, Greece was among the catching up countries showing scores lower than that of the average EU one, thus ranking in the last positions. This picture has evolved in the last year mainly favored by several statistical changes, as well stated both in the latest European Innovation Progress Report[88] as well as the 2009 Trendchart report, placing Greece among the moderate innovators category and giving a more positive reaction to its innovation system.

---

[88] European Commission, Inno Policy Trendchart, European Innovation Progress Report, 2008.

One of the major responsible bodies within the Greek innovation system dealing with STI is the General Secretary for Research and Technology (GSRT) which has been in force since 1985. Apart from its long existence, the most striking think is that over the years, it has undergone through frequent shifts in terms of responsibility being mainly a central agency for R&D with no substantial co-operation with sectoral ministries. This shift in responsibilities is quite evident until now that the GSRT is entrusted with innovation policy and manages approximately 1/3 of the overall public S&T budget[89]. Introducing an evaluation culture within the Greek innovation system is one of the values that the GSRT has brought to light even if it is in very limited terms. It is considered being the only public service that has actually organised a few evaluations of specific programmes and it has also introduced an evaluation of institutions tradition which has brought a very positive improvement in the supervised research centres[90].

Even if the public sector is being the main performer of innovation, a dualism has been identified which is mainly related to the role of ministries which relies on both the design and implementation of policies lacking at the same time of a specific public body that would therefore be responsible only for "*debating and agenda setting*"[91]. At the same time, GSRT who is mainly responsible for the STI activities in Greece has "*a very limited interaction with the sectoral ministries and agencies*" which makes it more difficult for the provision of qualitative results and successful policy making.

On the other hand, one of the weakest sectors in the national innovation system is that of business which is mainly composed of very small and traditional firms lacking a solid base of larger enterprises with strong R&D performance contributing to the slow adaptation of process innovation and an even slower development of a vast technological base. As well stated, "*Greek businesses are rather followers than innovation leaders and thus more involved in innovation transfer than innovation creation*"[92].

It was noted in the interviews that the Greek system is quite fragmented. The nature of this fragmentation lies on the absence of coordination among the different government entities relating to innovation. On the other side, stakeholders' involvement, referring to academia, is also limited to the policy design process whereas the private sector's limited involvement lies mainly in the formation of innovation policy due to the limited existence of modern and innovation driven firms[93].

Within the Greek innovation system and as far as complying with the SF regulations, "*a certification authority is set up whereas a management, auditing authority and follow up committee are so less active*". It has been acknowledged that the role of committees has been quite limited and fragmented. According to Tsipouri and Papadakou (2005), "*the agendas and frequency are more systematic but the emphasis is on speed and absorption at the expense of content*". Taking into account the SF regulations together with the standing situation of the Greek internal system, ministries do interact with each other in a certain way and only on the base of the different Operational Programme's monitoring committees. Auditing committees, being more structured, are mainly managing indirect controls whereas managing authorities are taking care of the implementation and supervision of the different OPs activities after their approval[94].

---

[89] Tsipouri, Papadakou, 2005

[90] Tsipouri, Papadakou, 2005

[91] Tsipouri, Papadakou, 2005

[92] GSRT, The Greek Innovation System, Review of Greece's Innovation Policy by the OECD, Background report, August 2007.

[93] N. Komninos and A. Tsamis, "The system of innovation in Greece: structural asymmetries and policy failure", *Int. J. Innovation Regional Development*, Vol. 1, No. 1, 2008

An accumulation of knowledge and expertise mechanism does not exist. The responsible directors are continuously changing whenever there is a change in the General or Special Secretary. In this regard and according to an interviewee, "*there is a very clear lack of know-how and a restriction and fading out of the transfer of knowledge*". According to Trendchart (2008), "*policy implementation suffers from deficient quantitative and qualitative staffing of the competent services, persisting failures in the design of procedures due to lack of evaluations, politics inside the civil service which leads to irrational management of programmes and human potential and hinders the accumulation of know-how and skills*".

With the creation of supervising authorities, for the different thematic areas, supposedly a continuation and monitoring of the overall innovation system was held but again these have mainly contributed to the further straining of the rest of the public administration's valuable "assets" (people). At this point, an interviewee stated that "*there is a constant lack of appropriate and qualified people but this mainly stems from the long lasting culture and quality of the public administration itself which leads to the inappropriate execution of evidence based policy making*". Adding to that, these authorities have again focused on the absorption and more or less mechanistic application of procedures and not on the identification and analysis of the results and impacts for the better exploitation of programmes and the efficient upgrade of the evaluation system.

4.1.2 Evaluation Culture

Project and programme evaluation was first introduced since the early 80s but they mainly concerned ex ante evaluations. After the introduction and involvement of Structural Fund policies, evaluation exercises were introduced in all stages of programme funding including interim as well as ex post evaluation. Even though they are introduced they do not require a very prominent position in terms of usage and application compared to the ex ante ones. Proper ex post evaluations are absent from the Greek innovation system posing a very important hindrance for the amelioration of the system per se as well as for the creation of clear policy lines. Even though such kind of evaluations were anticipated by the Operational Programmes for the period 2000-2006, their content was mainly focused on the absorption of funds rather than the actual results of the programmes as well as their impacts, weaknesses and opportunities for future amelioration[95].

According to Tsipouri, Papadakou (2005), "*policy learning remains limited and the use of modern evidence-based techniques is rudimentary; evaluation is marginal, for practical purposes it is limited to legal obligations via-a-vis the EU*". Given this fact, it is well accepted that such evaluations mainly lack reference to the programmes' overall impact as far as innovation is concerned. This issue is also regarded in the 2008 Greek Trendchart report where it was stated that "t*he evaluation as a mechanism of the Operational Programme has shown that, while the uptake and delivery systems of*

---

[94] Tsipouri, Papadakou, 2005

[95] Technopolis, 2006 in N. Komninos and A. Tsamis, "The system of innovation in Greece: structural asymmetries and policy failure", *Int. J. Innovation Regional Development*, Vol. 1, No. 1, 2008

each scheme are well investigated, there is a lack of impact assessment and appreciation of the scheme's qualitative achievements"[96]. In this regard and according to an interviewee, "*evaluations are conducted only for regulations to be delivered without giving any special attention to the results that could be obtained instead*". *"They are indeed conducted in order to comply with the European regulations set up for the smooth operation and absorption of Structural Funds. The results though could easily be characterised as being very gray."*

More specifically for each type of evaluation a brief description of the situation in the Greek Innovation System will follow indicating the most important issues that need to be covered. On the first instance and as far as the exante evaluations are concerned, the indicators produced are mainly used in strange ways. What the Greek evaluation system is missing is the basic idea behind these indicators: how do we evaluate and utilise these results? According to an interviewee, "*the basic problem also lies in the fact that there is a lack of primary sources for planning at the first place in order for them to be updated in every different programming period*". This issue leads us to a low or inexistent quality in the local evaluation system since the same sources are used all over again. Overall, we see that a practice such as assessment which has cultural roots in the functioning of society, is dissolved and degenerated in the Balkan Mediterranean, where there is a perception that politics is the supreme value.

Regarding mid term evaluations, it is said that "*more work has been done in this regard*". The problem though can mainly be attributed to the fact that this kind of evaluations happen at a very early stage without having acquired all the necessary elements. Coming from the same interviewee, "*we are obliged to conduct them so that we will be consistent with the SF regulations*". But this is neither the correct way to proceed nor the most prominent one for valuable results to be obtained.

The case with ex post evaluations in the local innovation system is very heavy (costive). Low level of planning as well as inexistent evaluation of previous measures and actions are the main issues that appear to have a negative impact on their effective functioning. Moreover, there is a serious quality problem in the programming mechanism and the programs per se. According to an interviewee, "t*he generality and polydisperse of the measures and actions under the respective Operational Programmes leave no room for serious assessment. Many actions are independent and should therefore be assessed separately".* In this way, apart from the absorption of funds, more concrete and clear results as well as the substantial impact of these measures in the local innovation system and evaluation culture will be obtained and further exploited.

Past experience in terms of the newly introduced measures in the new Operational Programme for Competitiveness and Entrepreneurship is highly needed and for this reason output results of the schemes as well as an analysis of their progress should take place[97]. It is widely acknowledged that "*a lack of evaluation of the previous schemes is a handicap in designing new measures as well as in improving the performance of the already established ones*"[98]. In this framework, it was noted in the interviews that the Greek system is very bureaucratic. There is little time for qualitative results. There is a huge lack of time and qualified people. In this regard, evaluations too formal and too general and thus conclusions are difficult to be drawn as far as additionality, effectiveness and impact issues of the individual measures are concerned, reducing the quality of inputs to new programmes and schemes[99]. Moreover, there is a prominent lack of funds for assessment and evaluations which constitutes another problem for the fulfilment of the Structural Fund evaluation requirements. According to an interviewee, "*these factors which are affecting the positive performance of a programme should be studied a priori*". In this regard and according to 2009 Trendchart, a new managing unit has been established within the GSRT which will be responsible in managing the measures of the OP Competitiveness and Entrepreneurship (2007-2013) as far as research and innovation are concerned in an effort to balance and enhance the development of newly established research and innovation measures[100] within the new programming period (2007-2013).

---

[96] Trendchart, 2008

[97] Trendchart, 2008

[98] Trendchart, 2008

The kind of evaluation culture that does exist in the Greek Innovation System mainly refers to innovation policy measures being evaluated at key milestones in their implementation and mainly when imposed by the European Commission[101]. It is of utmost importance the fact that the civil service system has not yet imbued a vast evaluation culture. There is a very limited use of evidence based design which is mainly attributed to the low amount of funds dedicated for evaluation. This deficiency does not allow for a more coherent impact assessment for the measures that have been applied and introduced so far. For instance, the interim evaluation of the Operational Programme for Competitiveness (2000-2006) mainly concentrated on the results and impacts of specific schemes whereas only two of them were aiming in promoting a more innovative character. One of the interviewees characteristically said that, "*public administration faces problems while trying to be consistent with the Structural Fund requirements*". Moreover, coming from the same interviewee, "*the first thing that applies is the values and the idiosyncrasy of the country, whilst Structural Fund requirements come second*". In this regard, there is a need for a balance between rigorous uniform regulation on the one hand and flexibility for the idiosyncrasies on the other. Despite the lower level of quality, the positive effects of the Structural Fund requirements in terms of building up awareness and capacity are overwhelming.

4.1.3 Conclusions

Careful steps should be taken into account in order to form a more coherent and effective innovation system. First and foremost, there is a need for more qualitative and continuous evaluation of the effectiveness and efficiency of the proposed measures. There is a need though for a more thorough and immense upgrading of the designing and managing authorities which are responsible for these measures. This upgrading should not only stand on developing and using all human resources but should also be based on continuous training in terms of project management issues and innovation systems development. According to Trendchart (2009), "*the most valuable good that may be produced and commercialized is new knowledge and that knowledge cannot be acquired in the same way as commodity goods".*

Concluding, a very good recommendation coming from the same report highlights the fact that "*building on existing strengths or building new strengths is an urgent task of any government*"[102].

---

[99] Trendchart, 2008

[100] Trendchart, 2009

[101] Tsipouri, Papadakou, 2005

[102] Trendchart, 2009

4.2 Malta

4.2.1 Innovation system

In the field of Innovation and Research Malta still seems to be in a primary stage, but considerable improvements have been made, since, despite the low innovation performance indicators, a steady and recently increasing growth rate is evident. As far as Malta's evaluation culture is concerned, it is still weak and underdevelopped, especially with regards to R&I. Currently there is no systematic scheme for policy-review in Malta's Research and Innovation Policy, but recent efforts have been made aiming to make it more formal, as a result of the country's recent EU membership.[103] Nonetheless, there is no standard format of evaluations in the country and no compliance with general Structural Funds (SF) requirements. However, when reviewing Malta's evaluation culture, one should take into account the nature of Malta's local innovation systems, which are subject to fast change and exhibit a rapid development rate. Hence, the requirements regarding evaluation design and execution must be treated with a greater degree of flexibility and is the reason why the country of Malta has not fully complied yet. The country is still going through its early stages of Research and Innovation and thus changes in project implementation require flexibility. [104] The lack of standardisation and the nature of evaluation in Malta will be further discussed in the following section.

4.2.2 Evaluation culture

As far as the key actors of the Maltese Innovation system are concerned, Malta's Key Evaluation Unit and Managing Authority for the implementation of major Operational Plans since 2001 has been the office of the Prime Minister – Planning and Priorities Coordination Division (PPCD)[105] – formerly known as the Regional Policy Directorate. The PPCD "*was set up by the PM[106] in 2001 as part of the administrative infrastructure required to manage the pre- and post-accession funds allocated to Malta by the EU*." The Planning and Priorities Coordination Division is the entity responsible for inter-agency co-operation at all stages of programme development and deployment including programming, monitoring, evaluation and reporting for Malta's Single Programming Document (SPD). Currently, the main operators dealing with research and innovation initiatives in Malta are: Malta Council for Science and Technology (MCST), Malta Enterprise (ME) and the University of Malta.

Evolution of the Maltese Evaluation Culture[107]

Through thorough analysis of Annual Innovation Policy Trends and Appraisal Reports on Malta, we deducted a number of conclusions and came across evidence clearly showing that the nature of the Evaluation Culture of the country up until now is still weak, at a developing stage and lacks systematisation. The main risk coming from insufficient evaluations and lack of in-depth analyses is that such problems may in fact, lead to the implementation of ineffective measures and poor planning, which seems to be the case for Malta at the moment, up to an extent. In 2005 an appraisal of the innovation governance system of the country it was clearly stated that Malta was missing formal policy making and evaluation practices in the fields of research and innovation, despite the fact that efforts were being made to introduce more systematic approaches to these procedures through better coordination and revision of the national framework for Research and Innovation. These reforms which took place in 2003 were implemented onthe Maltese RTDI policy–making cycle so as to make it more systematic. Lack of systematic approach was also evident on the policy review system and subsequently, this was reflected in the policy evaluation culture of the county in general. The Maltese culture of evaluations within the Government was poorly coordinated, weak in nature, , lacking systematisation as it was only starting to develop in the beginning of the present decade. This situation was a result of the lack of participation in EU programmes and the fact that evaluations up to that point were done on a small scale, as they were usually carried out by small panels of local experts with inputs from external experts, and their format depended entirely on the requirements of the respective clients..

---

[103] Cordis ERAWATCH National Profiles, Malta, http://cordis.europa.eu/erawatch, last accessed December 2009

[104] Based on Interview Template with local expert of evaluation and SF relevant procedures, Jennifer Harper, Director of Policy, Malta Council for Science and Technology

[105] http://www.ppcd.gov.mt/home?l=1, last accessed December 2009

[106] Prime Minister

[107] Based on Pro- INNO Trendchart: Annual Country Reports for Malta (2004 – 2008) - http://www.proinno-europe.eu, last accessed December 2009

Malta's evaluation culture did not seem to evolve or undergo any form of radical change in the year 2006, since there was still no formal mechanism or body responsible for the evaluation of Malta's innovation policy measures. Thus, new policy measures were drawn with the mere use of statistical data, as opposed to in depth analyses. Up until 2006, the most comprehensive evaluation carried out on a systematic manner and a base upon which future policies could actually be drawn was the European Innovation Scoreboard. Initiatives were taken though, which will be discussed in the following section.

The annual country reports for Malta, carried out both for 2007 and 2008, generate the same conclusion, i.e. that no significant progress was made and the rate of formal reviews and evaluations which were undertaken was low. The nature of evaluations carried out by the year 2007 depended on the guidelines of the respective Ministry or State Agency involved, but the general rule was the execution of internal evaluations usually carried out by independent experts. These reports remained internal and thus inaccessible by the public. As it was reported in 2005, the online form of the publications and evaluation results – as well as the content- depended on the nature of the published document, due to the fact that certain evaluations with potential impact on political issues were handled accordingly in order to ensure that sensitive matters would stay intact, mainly due to the highly politicised culture of Malta. Publicity of such documents however, does not substantially affect the evaluation culture of a country, as it seems to be one of the least important factors and definitely not a priority. An interviewee characteristically stated that "*although dissemination aspects are very important for evaluations, emphasis should be placed more on the content of the publications to be disseminated rather than ensuring that publicity will be given to the EU Support for a project*". [108]

More specifically and focusing on certain sectors and fields, based on data of the 2007 report, the evaluation culture in the field of innovation policy was still poor and there was room for considerable improvement, since the nature of innovation policy measures require constant evaluation and in a more systematic manner. After all key elements of evaluation procedures are continuity and good organisation and no such requirements were met. As for regional authorities, in the year 2007, the culture of evaluation of policies and measures seemed to be getting slowly incorporated into the local culture, as local authorities corresponded to the Government thorough their operational activities which were also audited by the Department of Finances. Keep in mind though that there is still no specific mechanism or body on regional level that reviews policies and none with regards to innovation. These poor evaluation practices, on all fields, are of course reflected on policy design, which needs much work in order to be carried out properly, since as we mentioned well planned evaluations can in fact lead to much higher quality plans and measures, as they will be based on evidence based tangible results.

---

[108] Based on Interview Template with local expert of evaluation and SF relevant procedures, Jennifer Harper, Director of Policy, Malta Council for Science and Technology

Thus, the general conclusion we can draw from going through all country reports of Malta, up until very recently (2009) the evaluation culture of the country still seems to remain unaffected and weak and with regards to publicity issues, in some public organisations and agencies evaluations actually remain unpublished and kept internal to the respective ministry or agency.

Public initiatives

In the past recent years, efforts have been made to strengthen and systematise the evaluation system of the country and the R&I system, either directly or indirectly.

Malta's National Strategic Reference Framework and the system of Research and Innovation

The year 2006 was crucial for Malta; despite the ongoing weak evaluation culture of the country, especially for innovation policy measures since key priorities were identified, a fact which actually made possible a better allocation of funding. A National Strategic Reference Framework (NSRF) was drawn based on a consultation process involving the participation of innovation players (both public and private), which was published by the Government. Nonetheless, *d*espite the uptake of such initiatives, there was still scope for cooperation and coordination among innovation players and a better coordination between public and private organisations as there are significant overlaps and so do the target groups. [109] No provisions were actually made for on-going assessment on the effectiveness of innovation measures on the NSRF, as it mainly promoted self-performed reviews for measuring progress, but it did provide a more systematic direction. As a result, the Research and Innovation System of the country was given a higher priority on the national policy agenda, a decision which was taken by taking into account key recommendations and proposals which resulted from older reviews.

Malta's Adherence to Structural Funds Regulations

According to Structural Funds (SF) regulations, which clearly specify the structures for SF interventions, there are a number of bodies which are currently responsible for a set of activities, namely a management authority, a certification authority, an auditing authority and lastly a follow up committee. The interviewee characterized this structure "*too focused on the financial aspects and correctness of structural funding and implementation*" and suggested that "*SF regulations should place more emphasis on content in terms of policy achievements and learning*". The main impacts from these SF structures were mostly related to "*greater financial rigor, ensuring that projects are managed according to set objectives and in a more timely way*", coming from the same interviewee who also confirmed that there has been lesser impact in terms of content evaluation and consequently on a practical level in Malta. [110]

---

[109] Cordis ERAWATCH National Profiles, Malta, http://cordis.europa.eu/erawatch, last accessed December 2009

The PPCD, namely the management authority actually created a system to better manage and allocate Structural Funds, the "Structural Funds Database" for the monitoring and implementation of the Single Programming Document (2004 – 2006), which was also created for the 2007 – 2013 period. The database of 2004 – 2006 was used for the maintenance of all the projects' details relevant to SF, maintenance of financial control of allocated funds, automated drawing up of reports and preparation of files to export data to the European Commission. The 2007-2013 SF Database as described by the PPCD is a "centralised system linked to the structural funds' stakeholders, which include the Managing Authority, Treasury, Certifying Authority, Audit Authority, Line Ministries, Intermediary Bodies and the Beneficiaries." 111

4.2.3 Conclusions

Although the creation of databases and initiatives taken form a substantial step forward and lead to better organisation of the Maltese evaluation system, emphasis is placed more on the typical aspects and financial data of projects; evidence that the country is still in need of a better evaluation culture, which needs much work to be done, especially with regards to the context of evaluations, which is a crucial factor for policy making procedures. Thus, the Maltese evaluation culture, still poor and underdeveloped, despite national efforts and attempts, has yet to be standardised and subsequently lead to a uniform pattern and scheme of evaluation procedures. This is also evident in the field of innovation-related programmes and policies, whereby there is no such thing as an "evaluation culture", but this should not come as a surprise, due to the fact that Malta has only recently joined the EU and as a result its innovation system is in an evolutionary level and still undergoing change. Hence, from a more realistic point of view, it would be fruitless to use a standard form of evaluation procedures and schemes on this stage, taking into account the primary conditions and currently evolving innovation system.

4.3 Poland

4.3.1 Innovation System

The polish National Innovation System has been described by its practitioners as "*fragmented and without adequate mechanisms of coordination between the key ministries responsible for innovation*"[112]. So far, the Economic Development Department of the Ministry of Economy as well as the Strategy and Science Department of the Ministry of Science and Higher Education are considered being the main mediums that are supporting innovation, science and technology policies but their coordination is lagging behind although there have been some sufficient steps towards the improvement of the situation.

According to Trendchart (2008), "*the Polish governance system is not optimal to foster innovation policies".* In this sense, specific changes should occur in order to avoid the foreseen challenges.

---

[110] Based on Interview Template with local expert of evaluation and SF relevant procedures, Jennifer Harper, Director of Policy, Malta Council for Science and Technology

[111] http://www.ppcd.gov.mt/home?l=1, Last accessed on January 2010

[112] Trendchart, 2008.

There is a need to "*reinforce the coordination of ongoing activities in the area of innovation policy by bringing innovation policy debate to the highest political level*". This reinforcement could be further endorsed with the "*creation of effective policy-making mechanism*". Furthermore, the application and functioning of a monitoring system which would supervise the implementation of the different programmes is of great importance. In this way both evaluation procedures and the overall control of the governance system itself will be enhanced[113].

So far and in the framework of the National Development Fund for 2004-2006 and the National Cohesion Strategy for 2007-2013, an Evaluation Studies Database has been set up where all evaluation proposals are in place. This tool which was developed upon request from the National Evaluation Unit of the Ministry of Regional Development supplies the public opinion with information with regards to the use made of EU funds in Poland and "*gathers and systematically organises all ex-ante and ongoing evaluation studies, whereby monitoring results are being stored*"[114]. As an interviewee characteristically stated, "*this centralised system has in many ways improved the quality of the evaluation system. Many researchers can easily find all information needed in order to check and monitor what has already been done*". It is a very efficient system whereby effective controls could be achieved but it is also an important source of information for conducting evaluation studies[115].

With the integration of the Polish Agency for Regional Development (PARD) into the structure of the newly established Polish Agency for Enterprise Development (PAED) a lot have been said about its existence which is under doubt and about the lost know how and valuable human capital over the years. PAED is though considered a very straightforward change within the Polish governance system and will be the responsible implementing authority of innovation policy measures within the new programming period (2007-2013)[116].

A major hindrance of the Maltese innovation system as far as research funding issues are concerned stems from the very weak science - industry cooperation which could thus be improved. This weak liaison is mainly attributed to the fragmented science system and the lack of a more competitive research funding. Businesses are reluctant in cooperating with science institutions either because it is not considered as a priority for them or because they simply never sought to establish such collaboration[117].

4.3.2 Evaluation culture

The evaluation system in Poland has undergone major changes in the last decade and so. Evaluation as a tool was firstly introduced and undertaken in the years 1993-1997 but with no substantial results as the system at that period could not support and provide its full attention to this subject due to the fact that there were no people with adequate skills and knowledge regarding the theory and methodology of the evaluation practices per se. The first time evaluation was actually put in the agenda was the years 1998-2001 whereby it became a major requirement of the country's regional policy. This could be attributed to the fact that the Polish administration undergone a major reform with the operation of new regulations, arrangements and the appearance of new actors acquiring different responsibilities within the Polish governance system. A major flaw in this respect was the fact that the five and most important elements of evaluation such as relevance, efficiency, effectiveness, utility and sustainability were not properly developed and explained which made evaluation less understandable and more confusing. The general guidance was insufficient and lacked specific and clear objectives and definitions[118].

---

[113] Trendchart, 2008

[114] http://www.ewaluacja.gov.pl/English/Research_findings/Strony/research_findings.aspx, last accessed 17 December 2009.

[115] Trendchart, 2008.

[116] Trendchart, 2008.

[117] Trendchart, 2008

In May 2004, when the accession period started, the Evaluation Unit was created. Under this framework evaluation of policies and programmes became essential starting firstly with the pre-accession funds and gradually focusing on evaluating actions which were co-financed by the structural funds. Evaluation was therefore used to evaluate the effectiveness and efficiency of public policies implementation. A preparatory phase of the evaluation process was executed in 2004-2005. At that time, the first evaluation researches started as well as the basics of evaluation capacity were being set up and developed. During the period 2004-2007, 135 evaluation researches were conducted of which 65 were ex ante evaluations[119].

The position of evaluation in the national innovation system has evolved over the years acquiring a more solid base. In 2004, evaluation was seen as a control instrument whereas in 2006 it was confronted as a legal obligation. From 2008 onwards, evaluation is seen as a management and accountability information source[120]. However, the evaluation culture of the Polish governance system needs to further undergo specific changes so that it could stimulate more interest with regard to innovation policies and programmes. According to the Polish trendchart report (2008), timing and budget are considered the major bottlenecks of the polish evaluation procedure due to the fact that too frequently evaluations do take place in a considerably immature period where recommendations cannot be reached and thoroughly examined. At the same time the unreasonable allocated budgets compared to the planned tasks make the evaluation process even more difficult and unmanageable[121]. This has mainly been observed during ex ante evaluations where funding was partial and inadequate so as to cover all aspects of the evaluation policy cycle.

Adding to this evolvement of the evaluation system in Poland, the financial resources attributed to evaluation conducted have increased from 100.000 € in 2004 to 6.250.000 € in 2008. The same goes for the evaluation reports that have been conducted over the last 5 years, resulting in five evaluation reports back in 2004 and reaching 112 reports in 2008. As for the personnel dedicated to the conduction of evaluations, in 2004 there were only 7 people reaching 30 in 2007 until this number has increased within a year reaching 153 people in 2008. These numbers are showing Poland's serious attitude towards the implementation of solid evaluations within their innovation system and their stable levels of development over the past 5 years[122].

---

[118] Karol Olejniczak, Towards the evaluation culture in Poland- experiences and prospects, Paper for the Annual Conference of regional Studies Association, 31st May and 1st June 2002.

[119] National Evaluation Unit, Department of Structural Policy Coordination, Ministry of Regional Development, "Evaluation Process of the cohesion policy in Poland- Summary of previous experience, plans and challenges for the future", April 2008.

[120] Bienias ppt, 2009

[121] Trendchart, Poland, 2008

[122] Bienias, ppt, 2009

Furthermore, the Polish administration has acquired a more upgraded knowledge on evaluation issues through the creation in 2004 of the Evaluation Steering Groups. These Groups were mainly created in order to decentralize the evaluation procedure and provide certain vigor to the interaction of the National Evaluation Units and the different external evaluation teams in all respective Operational Programmes[123]. According to Hoffman (2008), "*it was a long process of learning, but it was worth it, and our up-to-date experience with on-going evaluation is very positive*".

Capacity building in Poland

A very important phase of the evaluation process is the capacity building activities that the Polish innovation system aimed at further developing and enhancing on that period onwards. Institutions engaged in the evaluation process in the field of human resources development mainly hold the largest evaluation capacity in the whole system, especially the Managing Authority of the Sectoral Operational Programme Human Resources Development and the evaluation unit of the Polish Agency for Enterprise Development. Their capacity in the evaluation process "*stems from the competence of institutions' personnel as well as their managements' awareness of the significance of evaluation*"[124].

Acquiring an explicit capacity within the Polish administration is one of the major issues that have been pinpointed about the Poland's system. This not only serves for providing good evaluations but also forms a prerequisite for the transformation and further enhancement of managing the country's development policy[125]. Improving knowledge of the public administration in Poland is mainly acquired through several trainings organized by experts within the framework of the Twinning Programme as well as through specialized workshops co-financed by the Structural Funds within the framework of the Technical Assistance Operational Programme[126].

4.3.3 Conclusions

One of the major challenges the Polish evaluation system is facing corresponds to the fact that there is a need for better use of all the evaluation results as well as an improvement of the evaluation methodology not only in terms of impacts identified but also in terms of the evaluation's theoretical base per se[127]. The involvement and support of stakeholders in all stages of evaluation is immense for the future cooperation and implementation of recommendations. In this framework, one of the most important lessons that the Polish innovation system has learnt has to do with using evaluation as a tool for improvement, for future development of policies and programmes and for taking advantage of all its outcomes by providing solid and reliable recommendations for future amelioration[128]. The requirements and responsibilities of each of the EU Member States have evolved and changed over the years. As far as Poland is concerned and according to Hoffman (2008),

---

[123] http://www.funduszestrukturalne.gov.pl/English/Evaluation/NDP+CSF+20042006+Evaluation/ , last accessed 17 December 2009.

[124] National Evaluation Unit, Department of Structural Policy Coordination, Ministry of Regional Development, "Evaluation Process of the cohesion policy in Poland- Summary of previous experience, plans and challenges for the future", April 2008.

[125] Bienias, ppt, 2009

[126] http://www.funduszestrukturalne.gov.pl/English/Evaluation/Trainings+and+conferences/, last accessed 17 December 2009.

[127] Bienias, ppt, 2009

[128] Hoffman, 2008

"the role of evaluation has undergone a lot of changes reaching the point where it has become a significant factor in decision making process and which might have impact on national policies in a given time".

4.4 SF regulations: a need for institutional learning and structure building

This hypothesis was examined through the country cases. Indeed the results partly support the above statement based on positive impacts on capacity and structure building. On the other hand, institutional and policy learning and the establishment of sound evaluation systems still remains limited.

In Greece, despite the long-lasting experience in applying SF regulations, impacts from the created structures and required coordination have not managed to penetrate the system and overcome fragmentation among the key actors in the national innovation system. Policy learning has remained limited and the use of evidence-based techniques is marginal and limited to satisfying the obligations towards the EU. As an interviewee characteristically noted Greece is an example where, when policy implementation also suffers from deficient staffing, and unsound management, it is only possible to mechanistically abide by the rules and regulations rather than let them establish an evaluation culture and effective evaluation system. Thus, the impact of SF regulations remains limited.

However, there is wider acknowledgement that capacity building has been significant. At the same time, doubts were expressed about the suitability of SF regulations to lead to evaluations of high impact and usefulness to policy-makers. In particular, the lack of focus on impact assessment and evaluation of qualitative achievements was noted in this respect.

Malta is in a transition phase currently taking steps to respond to SF support requirements and regulations and establish an evaluation system. Interestingly, the points was echoed that the focus of SF regulations is much more on funds absorption and typical application of procedures rather than identification, analysis and real use of results and impacts for better exploitation of programmes and efficient upgrade of evaluation system.

Poland, on the other hand, is also a new Member State currently adapting to new requirements. Poland faces similar problems with Greece in terms of innovation governance (fragmentation, lack of adequate mechanisms for coordination between the key ministries responsible for innovation).

However, it has responded more effectively than Greece to the needs referring to an effective evaluation system. The requirements for monitoring the implementation of pre-accession funds resulted in creation of the Evaluation Unit. Since then evaluation received serious attention and was extended to SF supported actions. Evaluation staff has been significantly increasing while evaluation steering groups have also been formed to decentralize the evaluation system. Overall, the role of evaluation has undergone changes reaching now a point where it is a significant factor in decision making with potential impact on national policies.

Nevertheless, too much eagerness in implementation of the relevant regulations lead to the other end of the spectrum, i.e. too many evaluations in a still immature period and a difficult timing to examine and uptake the recommendations produced.

4.5 Explaining the survey results through the case studies

The findings from the examination of the country cases also provide some possible explanations for the survey findings presented in section 3 above.

The analysis of the situation in Greece reveals that SF evaluations are seen more as ‘internal’ exercises to the agencies initiating them and done merely for satisfying the obligations vis-à-vis the EU. SF regulations tend to be seen more as obligations that have to be typically met rather than as opportunity for essential institutional and policy learning. These elements may explain why the results of the SF evaluations are only to a limited degree discussed with government and stakeholders.

SF type evaluations in Greece seem to focus on the ‘letter rather than the essence of the law’. The conclusions produced may be too general in some cases while they do not necessarily address additionality and impact assessment, thus limiting the quality of inputs to new programmes and schemes. This can explain the limited usefulness of recommendations of SF type evaluations.

Other interviewees (from Malta too) echoed this point. It is doubtful whether SF regulations are indeed such that can lead to high quality evaluations with high impact on policy making. They are considered too much focused on financial aspects and correctness of Structural Funding and implementation rather than the real use of results and impacts for better exploitation of programmes and policy-learning. This can explain why even high quality SF evaluations may not lead to high impacts in terms of usefulness of recommendations and dissemination of results to national stakeholders.

At the same time, the variety of measures and programmes under the Operational Programmes calls for separate, more focused and specialized evaluations that cannot be accommodated under the SF regulations.

Notwithstanding that SF regulations may not be that fit to lead to high impact evaluations, there is also the fact that SF quality standards are only suggested rather than imposed. This, in combination with the typical application of procedures can explain why suggested quality criteria may not be applied in practice.

5 Conclusions

The evidence gathered in the INNO-Appraisal study shows that SF regulations provide guidance on how to conduct specific types of evaluations and this is reflected on which evaluation topics are addressed and data analysis methods are used in practice. Beyond this, the results reveal that despite that fact that certain quality criteria are suggested under SF regulations these are not necessarily followed in practice.

Secondly, evidence also shows that the specific quality criteria do not necessarily lead to high quality evaluations while the high quality evaluations do not necessarily have the highest impact in terms of usefulness of recommendations or discussions of results with government and stakeholders.

The country cases provide possible explanations for this phenomenon. The fragmentation among the key actors in the national innovation system in Greece, for example, and the fact that there is only typical abidance to SF regulations can explain why the results of SF evaluations are only to a limited degree discussed with government and wider stakeholders.

Abidance by the ‘letter rather than the essence of the law’ in combination with doubts about the suitability of the SF regulations to lead to high impact evaluations can explain the limited usefulness of recommendations as well as the fact that even high quality SF evaluations may not lead to high impacts in terms of usefulness and dissemination of results. The fact that SF regulations and quality standards are only suggested rather than imposed may explain why suggested quality criteria may not be applied in practice.

Finally the country cases show that the hypothesis that the demand of SF regulations in terms of high standards on structures and processes presupposes and leads to some institutional learning and structure building is partly confirmed. This is done based on the positive impacts reported on capacity and structure building. On the other hand, institutional and policy learning and the establishment of sound evaluation systems still remains limited.

References

1. CEC (2000), ‘The 2000-2006 Programming Period: The Mid Term Evaluation of Structural Fund Interventions. Methodological working papers, Working Paper n° 8’.
2. CEC (2005), ‘Communication from the Commission, Cohesion Policy in Support of Growth and Jobs: Community Strategic Guidelines, 2007-2013’ COM (2005) 0299, Brussels, 05.07.2005.
3. EC (2006a), ‘The New Programming Period 2007-2013. Indicative Guidelines On Evaluation Methods: Ex Ante Evaluation. Working Document No.1, DG Regional Policy, August 2006.
4. EC (2006b), ‘The New Programming Period 2007-2013. Indicative Guidelines On Evaluation Methods: Monitoring and Evaluation Indicators. Working Document No.2, DG Regional Policy, August 2006.
5. EC (2006c), ‘The New Programming Period 2007-2013. Guidance On The Methodology For Carrying Out, Cost-Benefit Analysis. Working Document No.4, DG Regional Policy, August 2006.
6. EC (2007), ‘The New Programming Period 2007-2013. Indicative Guidelines On Evaluation Methods: Evaluation During The Programming Period. Working Document No.5, DG Regional Policy, April 2007.
7. EC (1999a), ‘The New Programming period 2000-2006: The Ex-Ante Evaluation of the Structural Funds interventions. Methodological working papers. Working Paper 2’, DG Regional Policy and Cohesion.
8. EC (1999b), ‘The New Programming period 2000-2006: Indicators for Monitoring and Evaluation: An indicative methodology. Methodological working papers. Working Paper 3’, DG Regional Policy and Cohesion.
9. EU (2007), ‘Cohesion policy 2007–13 Commentaries and official texts’, Guide, Regional Policy, January 2007.
10. OJEC (1999), Council Regulation (EC) No 1260/1999 of 21 June 1999 laying down general provisions on the Structural Funds, (L161) June 1999.
11. OJEC (2006), Council Regulation (EC) No 1083/2006 of 11 July 2006 laying down general provisions on the European Regional Development Fund, the European Social Fund and the Cohesion Fund and repealing Regulation (EC) No 1260/1999, (L210), July 2006.
12. Smail, R. (2007), ‘Good Practice for Implementing Structural Funds Programmes and Projects’, EIPASCOPE 2007/3, pg. 13-19.
13. Tavistock Institute, GHK, IRS, (2003), ‘The Evaluation Of Socio-Economic Development. The GUIDE’, December 2003.

This page has been intentionally left blank.

Part III Chapter 8

Country Report: Austria

Michael Dinges, Julia Schmidmayer[129]

---

[129] The lead authors for this section of the report are as given above, with contributions from all other team members at different parts of its production, conclusions and interpretations have been approved and are shared by the group.

Table of Contents

Table of Contents 244
Table of Exhibits 244
Executive Summary 245
1 Innovation policies and evaluation practices in Austria 247
  1.1 RTI performance and innovation policies 247
  1.2 Evaluation practices 249
2 Stocktaking of data collection and analysis 250
  2.1 Policy Measure categorization 251
  2.2 Commission of evaluations 252
  2.3 Types of evaluations 253
    2.3.1 Topics covered in the appraisals 254
    2.3.2 Data collection and analysis methods 256
    2.3.3 Evaluation quality and usefulness 258
    2.3.4 Consequences of appraisals 260
  2.4 Summary, conclusions and lessons to be drawn 260
References 263

Table of Exhibits

Exhibit 69: Development of R&D expenditures as a percentage of gross domestic product by country 247
Exhibit 70: policy measure characterization 251
Exhibit 71: Tender procedure 253
Exhibit 72: Types of evaluations in Austria 2002-2007 253
Exhibit 73: Topics covered in the appraisal 255
Exhibit 74: Impact dimensions covered by appraisals 256
Exhibit 75: Data collection methods employed 257
Exhibit 76: Data analysis methods employed 258
Exhibit 77: Appropriateness of appraisal quality 259
Exhibit 78: Usefulness of recommendations 260
Exhibit 79: Timing and consequences of appraisals 260

Executive Summary

Being a laggard in terms of RTI investments until the mid-nineties, both public and private entities have increased R&D investments efforts tremendously in the last decade. Austria has exceeded the average R&D intensity level of the EU-15 and the OECD countries. But not only RTI funding has increased: Austria has a large stock of innovation promotion measures at hand: Apart from generous bottom-up RTI project funding schemes, a remarkable number of thematic R&D programmes, structural programmes, and tax incentives exist. Despite good overall conditions there are a series of systemic challenges that still need to be addressed (e.g. poor performance of the Austrian higher education system, insufficient framework conditions as regards regulations, poor private and public funding for innovative start-ups and spin offs).

During the catching-up process, RTI programmes were the most preferred way to address policy challenges. In this time, the use of evaluations increased dramatically. Evidence for the increased relevance of innovation policy evaluation is provided not only by evaluation counts, but by changes in the legal conditions for evaluations, measures to foster an evaluation culture, the transparency of evaluation results, and the high number of evaluation activities.

With 34 appraisal reports, Austria has the highest share of innovation appraisals in the Inno Appraisal database. Some distinct features of these evaluations are presented.

The majority of appraisals are carried out mid-term during one point in the programme’s lifetime. Mainly, a supportive purpose is followed as policy makers respectively programme managers need advice how to enhance programme implementation. Only a limited number of topics are addressed: Appraisals focus mainly on policy/ strategy development, output counts, and consistency matters. Whereas behavioural additionality issues are rather prominent in Austria, input and output additionality issues as well as quality of outputs are only considered in a limited number of evaluations. Technological, economic, and socio-economic impact dimensions are missing by large, or only refer to programme participants.

Low cost data gathering and data analysis methods prevail (descriptive statistics, context analyses, interviews, and monitoring data). Most commonly a mixed methodological approach where quantitative and qualitative methodologies are combined is used.

Compared with the other countries in the dataset, we see a significant lower coverage of input and output additionality issues, also the quality of outputs is widely neglected. Only a limited number of Austrian appraisals deal with impact at all: For every impact dimension coverage is lower in Austria than in the other countries of the dataset. If impact dimensions are covered they rather focus on direct impact than on the participants than beyond.

Partly, the low coverage of impact dimensions and certain topics might be due to the formative purpose of most evaluations. Another reason for the discrepancies is the high coverage of Austrian appraisals in the database. Whereas in Austria almost the full range of appraisals conducted in the field of innovation policy is covered, it is more likely that only bigger evaluations are covered in the other countries; significant differences as regards the tender procedure point in this direction.

Despite the intermingled picture as regards evaluation topics used, the quality of evaluations is perceived to be high by respondents. Given the evaluation purpose, also the methods used tend to be considered as appropriate. Especially, recommendations concerning changes to the management and implementation of RTI programmes were perceived to be useful. Forward-looking advice was regarded as helpful for the design and implementation of future policy measures.

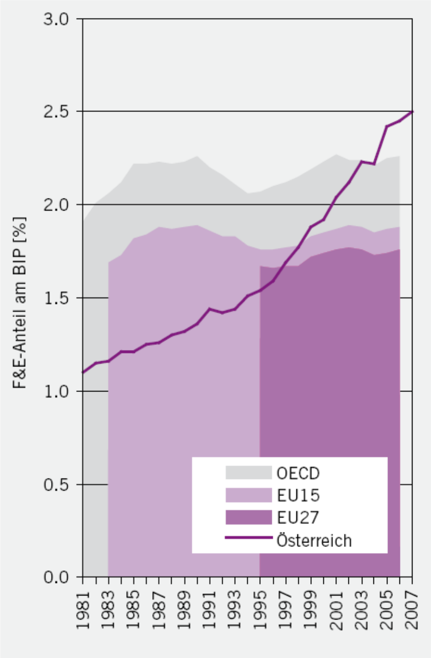
Nevertheless, due to the high number of evaluative activities, an increasing evaluation fatigue can be witnessed. Criticism was raised, that mechanisms ensuring that the results of evaluations do feed back into policy formulation and implementation are missing. In this respect, more thoughts need to be spent on the concrete purpose of planned evaluation activities, and the role of evaluations for policy implementation.

1 Innovation policies and evaluation practices in Austria

1.1 RTI performance and innovation policies

Being a laggard as regards R&D investments for a long time, and starting from clearly below average level of R&D intensity in the 1980s (1.1% as opposed to 1.64% of EU15 in 1981), Austria has undergone a period of almost twenty years in which its R&D intensity has constantly increased. In 1998 Austria has surpassed the R&D investment/GDP level of the EU-15 and since 2004 Austria's R&D intensity also exceeds the OECD average level (see exhibit below).

Exhibit 69: Development of R&D expenditures as a percentage of gross domestic product by country



	2000	2006	ø (2006–2000)
Belgium	1.97	1.83	- 0.14
Denmark	2.18 (1999)	2.43	+ 0.25
Germany	2.45	2.53	+ 0.08
Finland	3.34	3.45	+ 0.11
France	2.15	2.11	- 0.04
Greece	0.60 (1999)	0.57	- 0.03
United Kingdom	1.85	1.78	- 0.07
Ireland	1.12	1.32	+ 0.20
Italy	1.05	1.09 (2005)	+ 0.04
Netherlands	1.83	1.67	- 0.16
Norway	1.64 (1999)	1.52	- 0.12
Poland	0.64	0.56	- 0.08
Portugal	0.76	0.83	+ 0.07
Sweden	3.57 (1999)	3.73	+ 0.16
Slovak Republic	0.65	0.49	- 0.16
Spain	0.91	1.20	+ 0.29
Czech Republic	1.21	1.54	+ 0.33
Hungary	0.78	1.00	+ 0.22
Canada	1.92	1.94	+ 0.02
Japan	3.04	3.39	+ 0.35
Switzerland	2.53	2.90 (2003)	+ 0.37
USA	2.74	2.62	- 0.12
OECD	2.23	2.26	+ 0.03
China	0.90	1.42	+ 0.52

Source: OECD (MSTI), Calculations by Joanneum Research

Source: Austrian Research and Technology Report, 2009

In particular since the Austrian government committed itself to the Lisbon targets and the Barcelona process, as the national government in 2000 also aimed to increase R&D expenditures to 2.5% of GDP in 2006 and 3% in 2010 a series of efforts to spur innovation activities were launched. Main initiatives may be summarised as follows:

- **Introduction of new programmes** - Compared to other European countries a remarkable number of new policy measures were introduced in the field of innovation policy. In the last decade, about 60 RTI programmes were introduced to address structural weaknesses, foster science-industry relationships and close funding gaps.
- **Introduction of fiscal promotion measures -** Alongside the above mentioned programme-approach, also indirect, fiscal measures were introduced to leverage and promote R&D expenditures by firms.
- **Delegation towards agencies** - Facing the problem of a proliferation of RTI programmes the Industrial Research Promotion Fund (FFF) and the Austrian Science Fund (FWF) were evaluated in 2004 by an international consortium. The Austrian political actors implemented

the evaluation team's suggestions concerning governance structure which directly influenced the Industrial Research Funds' re-organisation. A completely new organisation was established, the so-called Austrian Research Promotion Agency (FFG). Continually, RTI programme management and implementation have moved from the three responsible ministries for research and innovation activities to the FFG. Furthermore a merger of economic service agencies and support agencies into the Austria Wirtschaftsservice Gmbh (AWS) (Austrian economic service) was accomplished.

Despite Austria's success story in terms of increasing its R&D intensity, which occurred mainly due to a tremendous increase of private R&D investments although government investments increased heavily as well, and an almost excellent position in the current (2008) European Innovation Scoreboard (EIS), which ranks Austria in sixth place at the top of the group of 'Innovation Followers', there are still quite some policy challenges that need to be addressed. Most prominent, the Austrian Council for Research and Technological Development, which is an advisory body to the Austrian Government, points in its new 2020 strategy in particular at,

- the below-average percentage of the population with a tertiary education qualification and the number of science and engineering graduates, and
- a comparatively low ratio in terms of the transformation of research input into output (i.e. Austria invests a disproportionally large share of resources in the RTI system and generates only lower than average output in comparison).

Apart from the poor performance of the Austrian education system and an average position as regards research output, a comparative cross-country study on innovation indicators (Hirschhausen et al. 2009) highlights at least three additional challenges (see also Schibany 2009), namely:

- Poor risk/venture capital funding and support measures for innovative start-ups;
- Poor framework conditions as regards regulations and poor status of competition;
- Poor societal innovation climate (openness, riskiness, tolerance, attitude towards science and technology).

Also the high number of R&D programs and the existence of a 'programme jungle' is still widely discussed: On the one hand there are still discussions about gaps in the R&D promotion portfolio, whereas on the other hand debates concerning the efficiency and impact of existing R&D promotion schemes, such as the tax allowance for R&D (introduced in 2002) and the R&D premium for firms arise. Additionality issues, windfall gains due to heavy reliance upon industrial R&D promotion measures open to all firms, and the interaction between direct and fiscal innovation policy measures present integral challenges for policy makers and evaluators of R&D policy.

Reasons for the proliferation of R&D programmes are manifold (for a discussion see for instance Dinges 2010, Pichler et al. 2008, Slipersaeter et al. 2007, Schibany and Jörg 2005). First of all, Austrian R&D project funding was for a long time dominated by two largely independent acting funding agencies: the Austrian Science Fund (FWF) and the Austrian Industrial research promotion fund (FFF), both founded in 1967. Both FFF and FWF were highly dependent on Government for their funding, but the internal governing structures and relationships with the beneficiaries have made them largely unresponsive to Government policies. As regards innovation policy, the legal mandate and mission of FFF provided an excellent basis for gaining wide responsibilities for innovation policy, but the agency restricted itself mainly to a narrow concept of bottom-up research-project funding by use of a traditional and well-established set of instruments acclaimed by the beneficiaries (Arnold, 2004). The lack of response first made Government initiate new agencies and a series of new instruments/programmes outside FFF to address long-lasting policy challenges such as fostering science industry co-operation, building critical masses in industrial R&D. In the end, this provoked a total re-organising of the councils by Government, introducing a much stronger and direct Government influence on them from 2004 onwards (Lepori et al 2007).

Apart from the policy challenges and the condition of the research funding agencies until 2004, also the Austrian governance structures for RTI policy contributed to a proliferation of program funding. In Austria three ministries share the competencies for RTI policy (Ministry of Science and Research, Ministry of Transport, Innovation and Technology and the Ministry of Economic Affairs). For the ministries RTI programmes constitute not only a good mean to address needs of the RTI system but also to safeguard and possibly expand their sphere of influence (cf. Schibany and Jörg 2005).

1.2 Evaluation practices

The catching up process of Austria’s RTI policy was accompanied by a series of activities which included e.g. delegation towards agencies, professionalization of allocation within funding agencies, programme planning, and evaluations. Policy makers and funding agencies used - or at least tried to use - evaluations as a tool to pursue a more rational policy approach. Hence, starting from the mid 1990s a certain evaluation culture in the Austrian RTI policy arose, which shows some distinct features:

- **Measures to enhance evaluation culture and transparency of evaluations:** Since its foundation in 1996 the Platform Research and Technology Policy Evaluation (fteval) pursues its mission “to encourage more, better and more transparent evaluations for an optimal strategic planning of RTD-policy in Austria and to develop a culture of evaluation”. Starting as a loose cooperation of people dealing with RTI programme design and policy analysis, the Platform is now described as - a highly institutionalised network, which comprises policy analysts, policy-makers and programme managers in the field of R&D policy (Edler 2007). The development of the platform fteval is marked by milestones such as
    - the publication of “Standards in Research and Technology Policy Evaluation” (2003), which provide both a framework and a set of guidelines for the evaluation process,
    - the organisation of two international evaluation conferences in Austria (2003, 2006) which provided an opportunity for evaluators and political stakeholders to discuss the current best practices and challenges for evaluations,
    - the compendium “Evaluation of Austrian Research and Technology Policies” (2007) - a summary of Austrian evaluation studies, and
    - the provision of a series of training courses for agencies and ministries in which evaluation methods and its potential use and limits are presented.

  The activities of the platform also contribute to an overall high transparency of evaluations in the field of RTI policy as the platform a) keeps an up-to-date database of evaluations performed, b) organises events in which recent evaluations and methodological issues are presented, and c) edits a newsletter in order to contribute to methodological discussions, challenges in evaluations, and the dissemination of evaluation results.

- **Uptake into RTI policy recommendations and legislation:** The evaluation activities in Austrian RTI policy also became a more binding control mechanism as the legal condition for evaluations in Austria were to some extent formalized. In 2005 the Austrian Council for Research and Technology Development recommended the active implementation of evaluation as an instrument in policy making. A further step involved the directives for the advancement of economic-technological research and technology development in 2006. These so-called RTDI directives (FTE-Richtlinien) state that “a written evaluation concept must be provided, containing the goal, the aims, and the procedures, as well as the dates for controlling the achievement of the advancement aims for all advancement programmes that are based on the FTE directives”. It also calls for the implementation of monitoring procedures.
- **High Frequency of evaluations:** Between 2003 and 2007 more than 50 RTI policy evaluations were carried out (see Zinöcker 2007). This is quite a lot for a country as small as Austria. Most obvious, reasons for this are the high number of existing RTI programmes in Austria and the formalisation of evaluation requirements.

Overall, the evaluation conduct in Austria in recent years shows that, evaluations constitute an integral part of the Austrian RTI policy system. Zinöcker (2007) points out that (some) decision makers in the agencies and in the ministries have seriously taken heed of evaluation results, leading them either to accept some recommendations or (justifiably) eliminate programmes. However, the high number of evaluation activities also led to a rising evaluation fatigue recently. As most of the evaluations dealt with single programmes only, single measure evaluations did not solve the issue of the existence of a high number of policy measures. This raises the question how results and recommendations of evaluations may better feed back into policy formulation and programme implementation (see CREST Policy Mix Expert Group: Country Report Austria, 2008).

While the Austrian ministries in 2008/2009 risked an ambitious attempt to move towards a system evaluation which should give an insight into the Austrian policy system and the interaction of R&D programmes, it is still not clear whether changes in the innovation promotion system will take place. After the results of the system evaluation have been published the Austrian government launched a process for developing a RTI strategy. The strategy process is expected to be finalised by June 2010.

2 Stocktaking of data collection and analysis

In the INNO-Appraisal dataset, which rests upon the Trendchart database, Austria is the leading country in the sample according to the number of appraisals. It has 34 single appraisals for innovation policy measures as listed in Trendchart, in total there are 171 appraisals (portfolio evaluations are only counted once), the second largest country group is Germany with 18 appraisals. 1 out of the 34 Austrian appraisals is a portfolio evaluation dealing with more than one policy measure. In order to avoid bias in the statistical analyses, all portfolio evaluations are only considered once in the dataset.

The above average representation auf Austria in the dataset is not only due to the high number of policy measures, but also due to the activities of the Platform Research and Technology Policy Evaluation (fteval), which has published a compendium “Evaluation of Austrian Research and Technology Policies” (2007), containing a summary of Austrian evaluation studies. Furthermore the

Platform contains an up to date database in which evaluation results of RTDI policy measures are made publicly available. As a consequence, availability of appraisal reports was much better than in other countries.

2.1 Policy Measure categorization

The Inno Appraisal database characterizes policy measures according to their modalities and target groups. This characterization is based on categories employed in TrendChart. The number of modalities was reduced by the Inno Appraisal team in order to minimize multiple response coding to a maximum of three modalities. The exhibit below displays the characterization of the Austrian policy measures vs. the rest of the dataset. Most of the evaluated measures combine two or three modalities. Overall, only minor differences as regards types of policy measures can be witnessed.

Exhibit 70: policy measure characterization

	Frequency in AT		Frequency in other countries	
Modality of the evaluated Policy Measure	Count	%	Count	%
Indirect measures (M1)	0	0%	7	6%
Direct financial support (M2)	17	53%	71	61%
Non-R&D related support (M3)	9	28%	36	31%
Creation of intermediary bodies (M4)	3	9%	8	7%
Mobility of Personnel (M5)	1	3%	12	10%
Creation of start-ups (M6)	2	6%	11	9%
Networks & Clusters (M7)	2	6%	43	37%
Science-industry cooperation (M8)	6	19%	31	26%
Support for the uptake and diffusion of innovation (M9)	11	34%	21	18%
Target group of the Policy Measure	AT	%	ROW	%
Uni/ PRO	21	66%	81	69%
All Firms	19	59%	67	57%
SMEs only	5	16%	29	25%
Sectors	3	9%	26	22%
Regions	2	6%	21	18%
Other	1	3%	25	21%
Number of valid cases (evaluations)	32		117	

In Austria as well as in the other countries of the sample the most prominent way of innovation funding is via direct financial support. While the number of measures that deal with mobility of personnel, creation of start-ups and the creation of intermediary bodies show relatively low numbers for all countries, innovation promotion via networks and clusters is distinctly low in Austria compared to the rest of the sample. Due to timing of data collection for the Inno-Appraisal activity the table still does not take into account the evaluation of tax support measures in Austria, which took place as a part of the System-Evaluation (Aiginger et al. 2009)

Also as regards the target group of the measure only minor differences occur. Compared to the other countries in the sample Austria only has a limited number of sector/technology specific initiatives. Most of Austrian funding measures target at firms and universities. Although the Austrian economy is predominantly run by SMEs, the majority of funding measures is not restricted to SMEs only but open to all firms.

2.2 Commission of evaluations

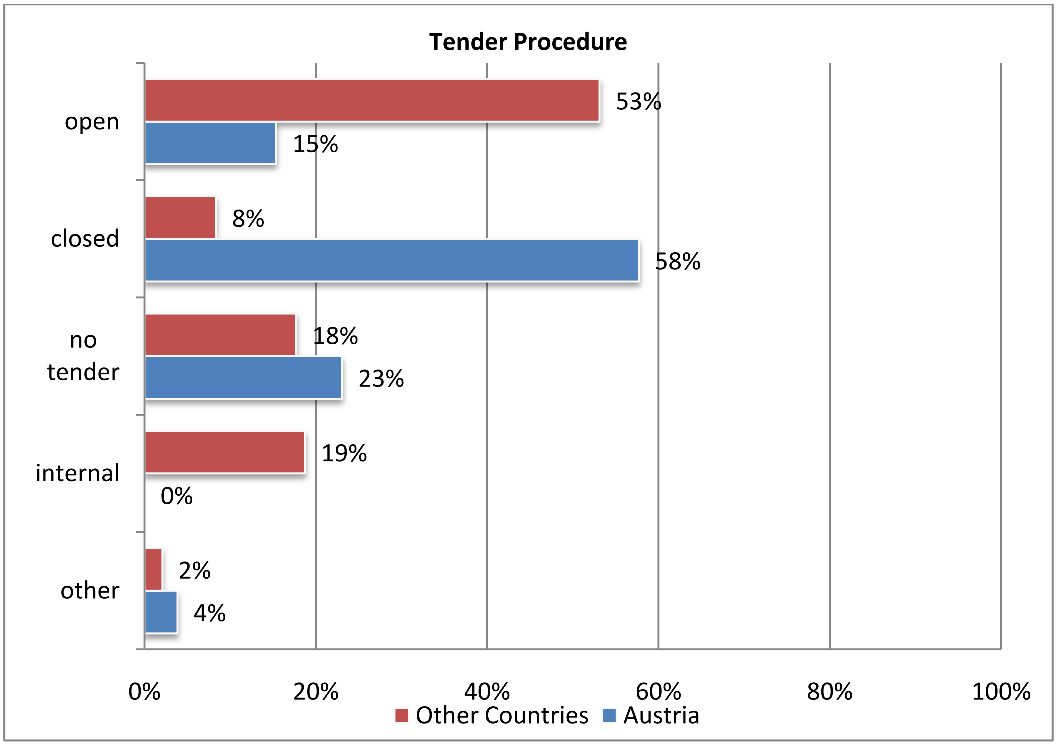
By large, the Inno Appraisal database mirrors the general knowledge about evaluation practices in Austria:

- The vast majority of evaluations were conducted by external evaluators (31). There are no internal evaluations that have been made public. Three evaluations took a mixed approach.
- Most of the appraisals (74%) were already foreseen and planned for during the design phase of the measure. This number is most likely to increase in the future as since 2006 the above mentioned RTDI directives make the planning of an evaluation concept obligatory for all new programmes.
- Most evaluations (53%) are commissioned during the lifetime of a programme and either accompanying or interim evaluations.

As regards the tender procedure a significant difference compared with other countries in the dataset is that most tenders were allocated through a “closed” tender procedure, which means that only a limited number of institutions are invited to tender: While in the overall dataset 53% of evaluations are conducted in an open procedure, only 15% of Austrian evaluation tenders use an open procedure (see Exhibit 71).

The preferred tender procedure is also reflected in terms of funding provided for the evaluation: Funding of appraisals covered in the database ranges between 10k Euros and 120k Euros. On average the level of funding is very moderate (55k). However, this seems to correspond firstly with an overall supportive character of evaluations and secondly, with the high number of policy measures in Austria, of which a good share has only limited availability of funds (see Dinges 2010).

Exhibit 71: Tender procedure



2.3 Types of evaluations

Among the numerous variables the Inno Appraisal database collects for each appraisal, the aspects of timing and purpose classify an evaluation study rather good. The project distinguishes among four types of timing, ex-ante, ex-post, interim and accompanying. Accompanying evaluations take place at several points in time, while an interim evaluation takes place at a single moment during the runtime of a program (e.g. at the end of a budget period). Ex-post is defined as taking place after the termination of the evaluated measure. Aside from timing also the purpose of evaluations was traced. In most of the analysed appraisals there is a blend of formative and summative purposes, however, one element is often dominant and therefore chosen, when coding.

Exhibit 72: Types of evaluations in Austria 2002-2007

Purpose Timing	*summative*	*formative*	*both*	*other*	*Total*
ex ante	0	5	0	0	5
accompanying	1	2	1	0	4
interim	4	10	4	0	18
ex post	5	0	0	0	5
other	0	1	0	1	2
Total	10	18	5	1	34

The cross tabulation above shows a total of 5 ex ante evaluations in Austria. Often, ex ante evaluations are conducted within the responsible ministry and hence not available for the public.

Nevertheless there also exist studies by external experts and activities within ministries and funding agencies that contain elements of an *ex ante* evaluation. These studies often go under the guise of "feasibility studies" introduced before, or at the beginning of new initiatives.

The vast majority of evaluations in the Austrian dataset of Inno Appraisal are interim evaluations (18). The interim evaluations mainly tend to serve a supportive (formative) role aimed at enhancing or readjusting programmes. The demand for interim evaluations has risen dramatically over the last years. Partially, this may be explained by the above mentioned recommendations of the Austrian Council for Science and Technology, but ministries also need to justify use of funds for their programmes. In particular, the ministry of finance (BMF) plays a distinct role in this respect as it decides upon allocation of funds even for single programmes.

Despite an increasing demand in terms of accountability and justification of money spent there are relatively few clear examples of *ex post* evaluations and impact assessments. All ex post evaluations in the sample have a summative character. The low number of ex post assessments might be due to the fact that these types of evaluation may only be relevant for policy makers as regards the design of follow-up programmes and are not relevant for important ad-hoc funding decisions.

Overall, the majority of evaluations in Austria have a formative character, especially interim evaluations are designed to give advice how programme management can be enhanced or readjusted. Summative elements in interim evaluations exist, but they mainly address the rationale of a programme and the priority setting within a programme. As programme managers are interested in setting a particular course in the "here and now", this type of information is of course the most relevant one.

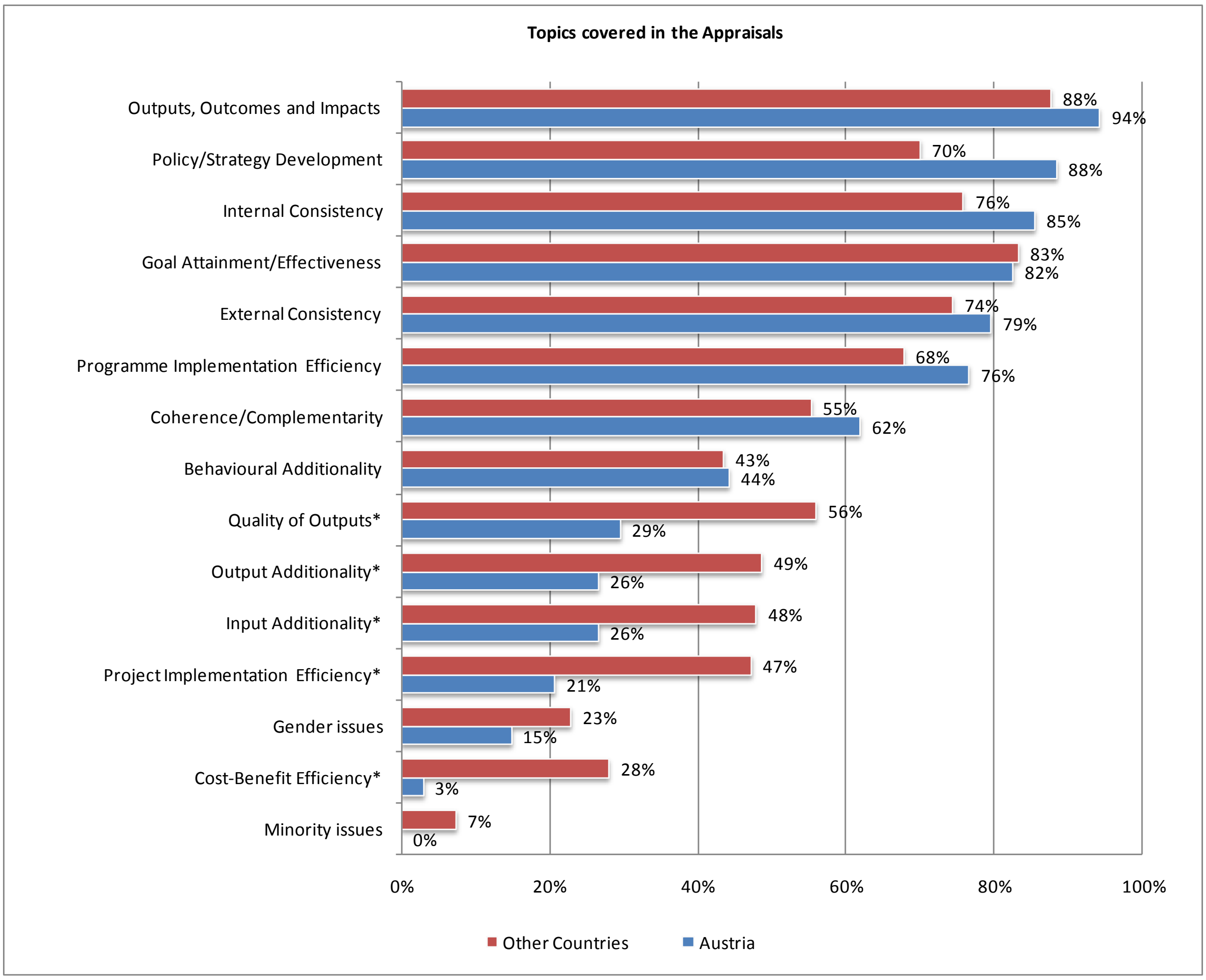
Although many programmes are evaluated at a certain point in a programme's lifetime, in most cases insufficient time has passed to conduct a full range impact assessment including economic and socio-economic impacts. Hence, only a small number of Austrian evaluations follow a summative purpose in order to judge the impact of a programme. The summative interim evaluations are mainly commissioned at an advanced point in the programme's lifetime.

2.3.1 Topics covered in the appraisals

Exhibit 73 provides an overview about topics covered in the appraisals. It highlights some significant differences between appraisals conducted in Austria vs. appraisal conducted in the other countries of the sample.

The most prominent topics covered in Austria are Outputs/Outcomes and Impact, Policy/Strategy development and Internal consistency. These issues also rank high in the appraisals of the other countries. Also behavioural additionality issues are rather prominent in Austria and the other countries. For Austria, this result might be attributed to the high number of interim evaluations which have by large a formative approach; changes in the behaviour of participants can be traced rather early in the lifetime of a programme by means of surveys and also qualitative methodological approaches. As most interim evaluations took place only some years have passed since the start of a programme, behavioural additionality issues are often the only impact dimension which can be covered at that point.

Exhibit 73: Topics covered in the appraisal



* The Chi-square statistic is significant at the 0.05 level.

Significant differences between Austria and the other countries in the dataset can be witnessed as regards topics that deal with the economic and technological impact of a programme.

- Only about one quarter of the Austrian appraisals cover input and output additionality issues whereas almost 50% of the appraisals in the other countries focus on these topics.
- Only 30% of Austrian appraisals cover the quality of outputs produced whereas 56% of appraisals in the other countries deal with these issues.

The low rate of input and output additionality measurement seems to correspond with the purpose of most of the evaluations (formative/interim). In such a context questions of input and output additionality or impact assessments are a) not top priorities and b) would need to make use of a sophisticated set of quantitative and qualitative methods which also requires heavy resources in time and cost. Also the low share in terms an assessment of the “quality of output” might be due to the by large formative character of the appraisals.

Taking into account the impact variables in the database, we witness that only a limited number of Austrian appraisals deal with impact at all. For every impact dimension coverage is lower in Austria than in the other countries of the dataset. If impact dimensions are covered in Austria, they rather focus on direct impact on the participants than beyond. Hence, although some basic output counts are provided in the evaluations, mid-term outcomes and impacts tend to be neglected.

Exhibit 74: Impact dimensions covered by appraisals

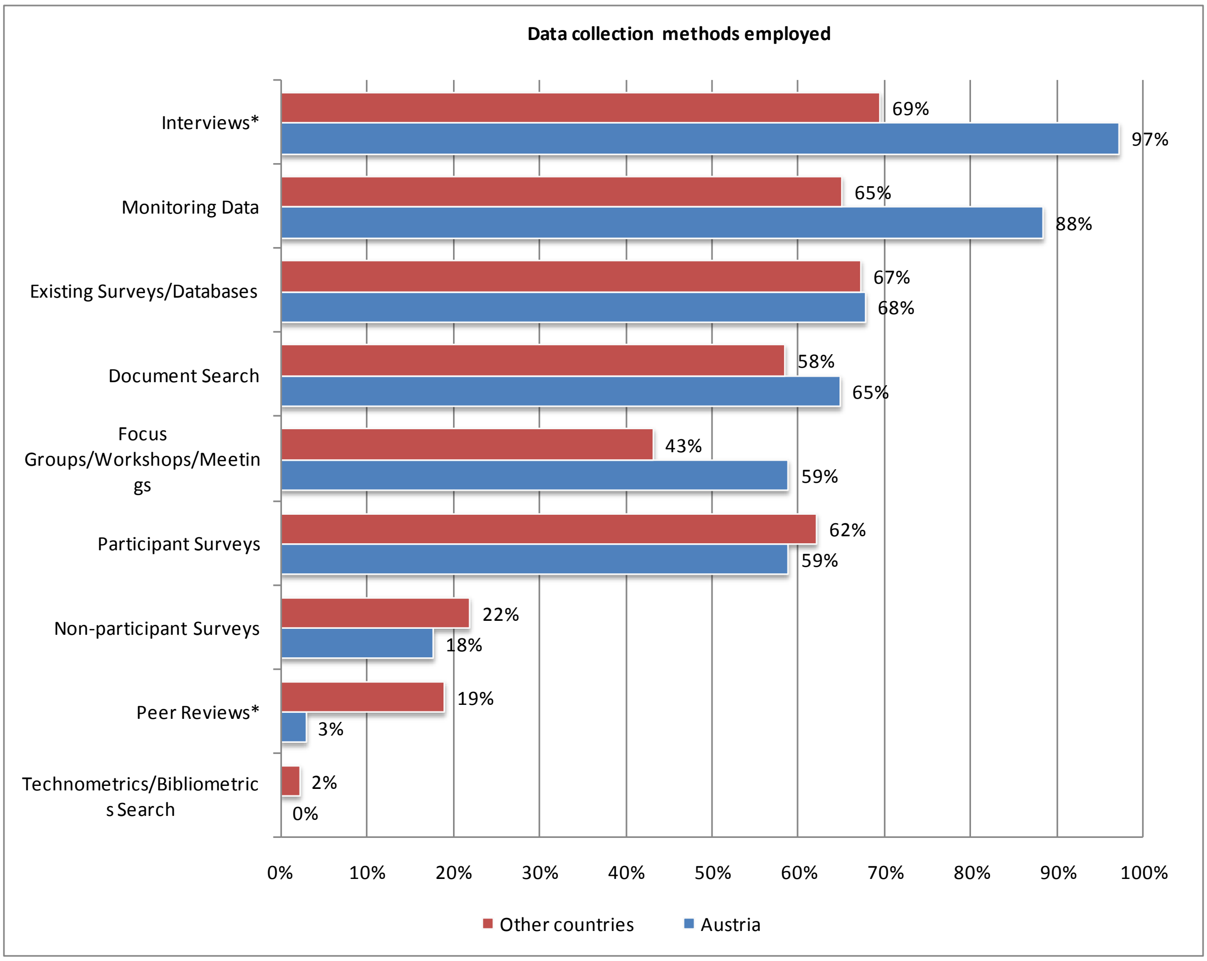
		Austria		Other Countries	
Scientific	Only on particpants	4	12%	19	14%
	Beyond participants	3	9%	41	30%
Technological	Only on particpants	6	18%	32	23%
	Beyond participants	3	9%	54	39%
Economic	Only on particpants	8	24%	36	26%
	Beyond participants	4	12%	78	57%
Social	Only on particpants	0	0%	8	6%
	Beyond participants	1	3%	60	44%
Environmental	Only on particpants	2	6%	5	4%
	Beyond participants	0	0%	34	25%

2.3.2 Data collection and analysis methods

Looking at data collection methods employed there seems to be a core set of quantitative and qualitative collection approaches in Austria. Interviews, monitoring data and existing surveys and databases constitute the most frequent data collection methods in Austria. Also document searches, focus groups and participant surveys are conducted in more than every second appraisal conducted in Austria. Interestingly, non-participant surveys, which might be necessary for conducting control-group approaches in evaluations are only used in about one fifth of the evaluations in Austria and internationally.

In addition, the international comparison shows that only some minor differences between Austria and the other countries in the sample exist. Austrian appraisals make significantly more often use of interviews, but the overall level of use of interviews is high. Peer reviews in programme evaluations are almost completely absent in Austria, whereas about one fifth of international evaluations make use of peer reviews.

Exhibit 75: Data collection methods employed

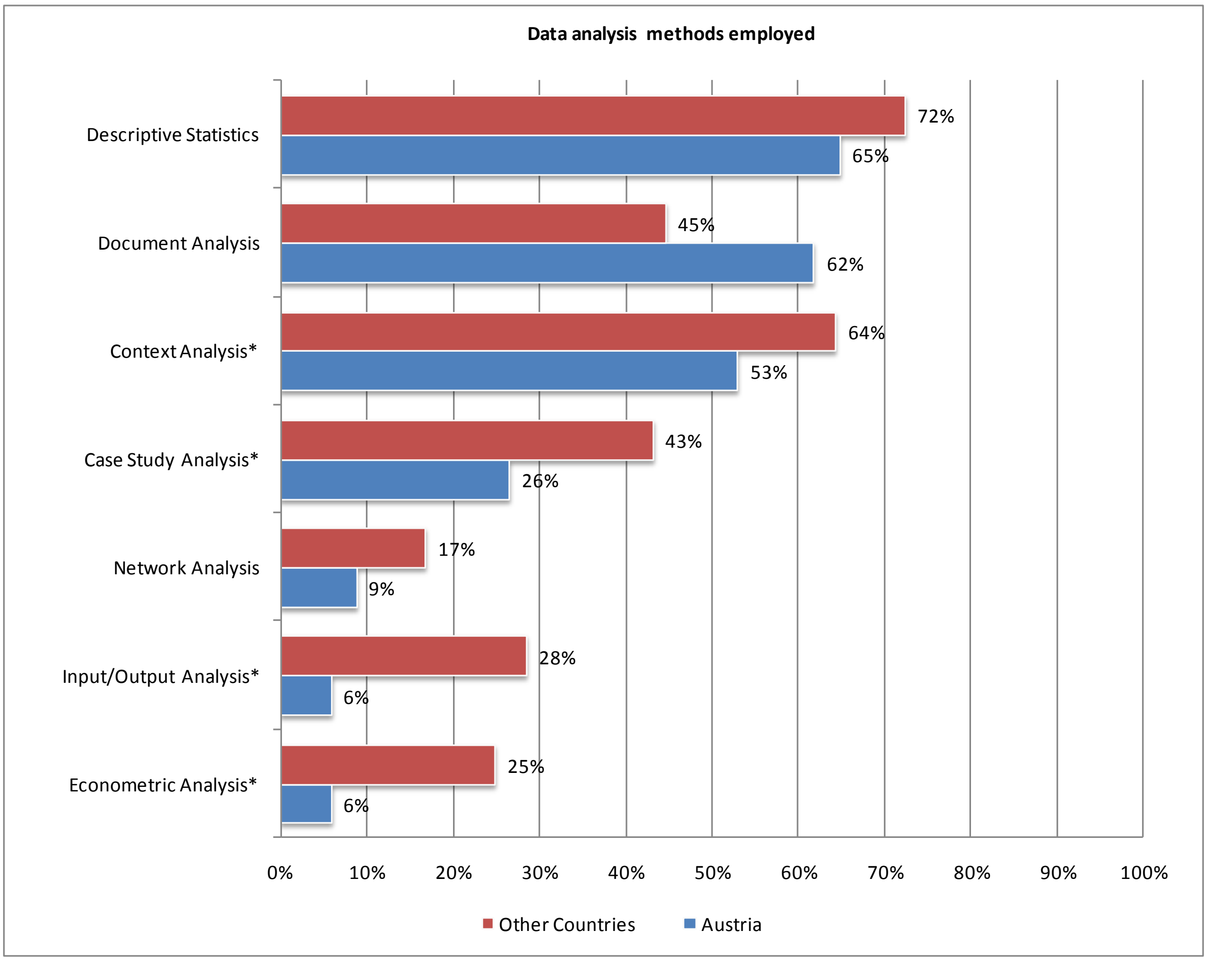


* The Chi-square statistic is significant at the 0.05 level.

Regarding data analysis methods the majority of evaluations apply a mixed methodological approach where quantitative and qualitative methodologies are combined. Some point at strong methodological developments in the last years as Logic Charts, Logit/Probit Analysis, Matched Pairs Analysis, Network Analysis and Focus Groups were introduced for the first time (cf. Zinöcker 2007, Zinöcker and Dinges 2009). However, when considering the total of evaluations performed in Austria we see that simple descriptive statistics, document analysis and context analysis build the core of data analysis methods. More sophisticated quantitative methods (econometric analysis, control group approaches, network analysis) are used only in very specific cases.

The international comparison shows again that some interesting differences between appraisals in Austria and the other countries exist. In line with the coverage of topics, Input and Output analysis as well as econometric analyses are significantly used less often in Austria, although in other countries these methods are also not employed in a large number of evaluations. But not only these quantitative methods are used less often in Austria. Also case study analysis and context analysis are used less often in Austria than in other countries. This is somewhat surprising as Austria has a high number of formative interim evaluations. In this setting, context analyses could, for instance, provide valid information as regards a programme's rationale. Case study analyses could be used to cover for instance impact dimensions and issues of programme implementation efficiency in a qualitative manner.

Exhibit 76: Data analysis methods employed



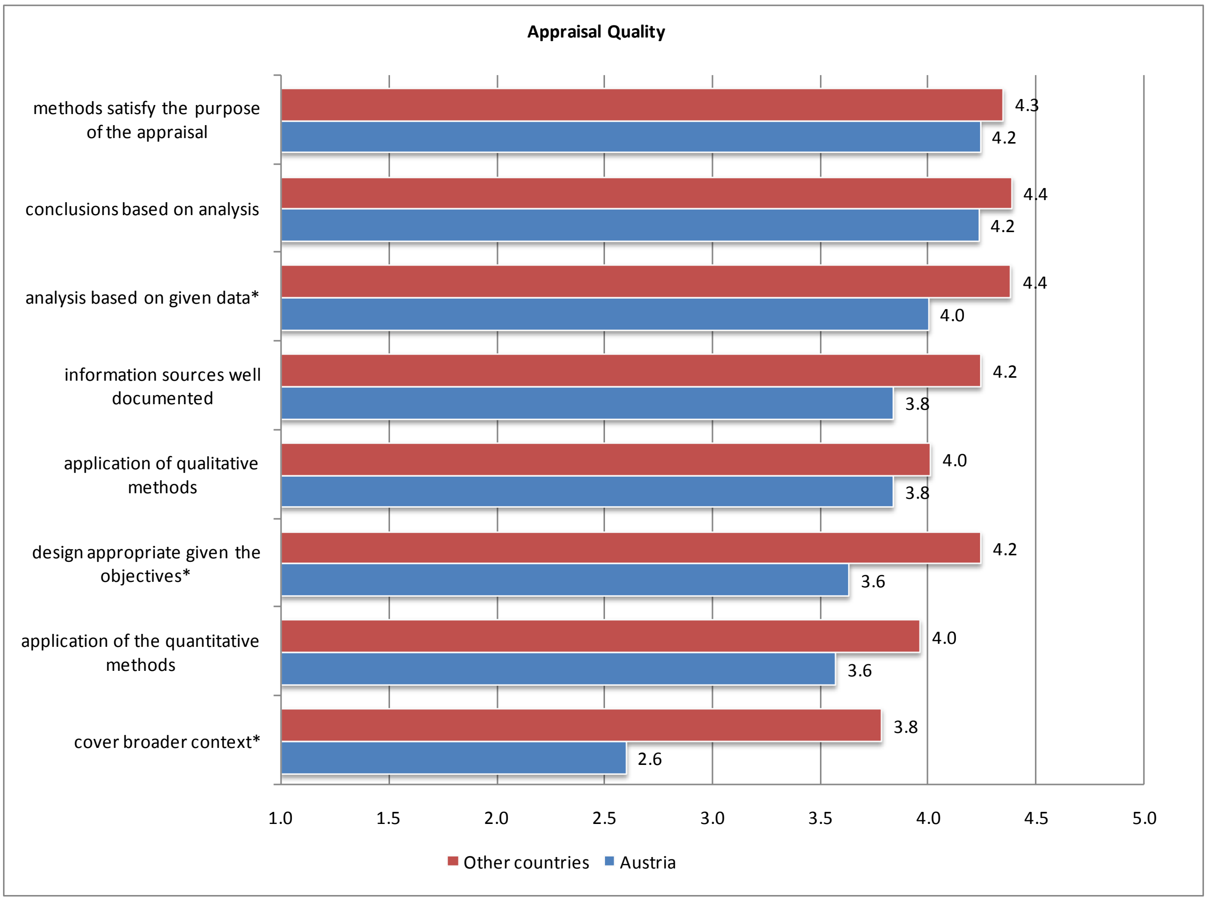
* The Chi-square statistic is significant at the 0.05 level.

Overall, the relatively low level of funding and the formative character of Austrian appraisals seems to influence the choice of topics covered and methodologies used. Questions of additionality or impact assessments need to be scrutinised and bring along a sophisticated set of quantitative and qualitative methods which demands more efforts in time and cost.

2.3.3 Evaluation quality and usefulness

The quality of appraisals was ranked by policy makers in the range between 1 (no, not at all) and 5 (yes, definitely). Despite the heavy reliance upon descriptive statistics, document analysis and interviews respondents provided rather high rankings for the different quality aspects covered in the Inno Appraisal questionnaire.

Exhibit 77: Appropriateness of appraisal quality



* t-test for Equality of Means significant at 5% level

Given the evaluation purpose, the methods employed tend to be considered as appropriate. Evaluation teams seem to choose a balanced set of methodologies which satisfy the purpose of the appraisal. The majority of appraisals focus on the usage of qualitative methods. Thus, the application of quantitative methods is not as satisfactory as of qualitative ones. The coverage of the broader context is not considered extensively in the Austrian evaluation reports rated.

As regards the usefulness of recommendations, policy makers had a bit less optimistic perspective: While recommendations concerning changes to the management and implementation of programmes were perceived to be useful and forward-looking advice was regarded as helpful for the design and implementation of future policy measures, recommendations regarding changes in the design of the measure and to other contemporaneous programmes only show average ratings. In this respect an important criticism was raised by the CREST policy mix review team "more thought should therefore be given to the mechanisms needed to ensure that the results of evaluations do feed back into policy formulation and implementation" (see CREST Policy Mix Expert Group: Country Report Austria, 2008, p. 17). Although, recommendations of appraisals were discussed within different interest groups, mechanisms are missing to ensure their sustainability.

Exhibit 78: Usefulness of recommendations

	Austria	Other countries
Internal Usefulness		
design*	2.6	3.3
management/implementation	3.3	3.3
External usefulness		
design/management/implementation of future programmes/measures	3.4	3.6
design/management/implementation of contemporaneous programmes/measures	2.2	2.3
broader policy formulation and implementation	3.0	2.9

* t-test for Equality of Means significant at 5% level

2.3.4 Consequences of appraisals

Contrary to the CREST expert team's criticism some recommendations of evaluations were implemented and caused at least a minor re-design of policy measures. The 19 interim evaluations cover the full range cover the full range of possible consequences: 2 programmes were terminated, 3 programmes had to undergo a major redesign, and in 8 cases minor re-designs occurred. The few number of ex ante evaluations were used to justify setting up RTI programmes and helped to adjust the design of policy measures being planned. Interestingly, some evaluations also had an impact of another programme – which is an evidence that the funding environment was considered at least in some cases.

Exhibit 79: Timing and consequences of appraisals

	Termination	Major Re-design	Minor Re-design	Expansion	Re-design of other measure	Total
	Nr (%)	Nr (%)	Nr (%)	Nr (%)	Nr (%)	Nr (%)
ex-ante	0 (0%)	1 (25%)	1 (25%)	1 (25%)	1 (25%)	4 (13%)
Accomp.	0 (0%)	0 (0%)	3 (75%)	1 (25%)	0 (0%)	4 (13%)
interim	2 (11%)	3 (16%)	8 (42%)	4 (21%)	2 (11%)	19 (61%)
ex-post	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (100%)	1 (3%)
other	0 (0%)	1 (33%)	1 (33%)	1 (33%)	0 (0%)	3 (10%)
Total	2 (6%)	5 (16%)	13 (42%)	7 (23%)	4 (13%)	31 (100%)

2.4 Summary, conclusions and lessons to be drawn

For the Austrian innovation system, the last decade can be characterised as a period of catching up and change. Being a laggard in terms of RTI investments until the mid-nineties, both public and private entities have increased their efforts tremendously. In terms of R&D intensity Austria nowadays has finally exceeded the average level of the EU-15 and the OECD countries.

Also in terms of policy instruments Austria has undergone a process of change and catching-up since the mid 1990s. Apart from a generous bottom-up R&D project funding scheme for firms and individual scientists, Austria nowadays shows a remarkable number of direct policy measures to address structural weaknesses, foster science-industry relationships and close funding gaps. In fact, due to distinct conditions at the governance level - 3 ministries responsible for RTI and largely autonomous funding agencies reluctant to change their scope of activities until 2004 - the most preferred intervention mechanism of policy makers were new RTI programmes. In addition, tax incentives open to all firms were introduced to further subsidize R&D in firms.

Overall, in terms of RTI promotion measures, the Austrian system can nowadays be characterised as generous and by far complete. Despite these overall good conditions there are a series of systemic challenges that need to be addressed. In particular, Austria suffers from a poor performance of the Austrian higher education system, which constitutes a real bottleneck to further increase of innovation activities. Furthermore, insufficient framework conditions as regards regulations, poor private and public funding for innovative start-ups and spin offs, and last but not least a poor societal innovation climate exist.

During the catching-up process described above, we witnessed that the conduct of evaluation for certain policy measures got increasingly binding, and that the number of evaluations performed in the last decade has increased tremendously. In this respect, the Inno Appraisal activity highlights some distinct features of the recent evaluation practice in Austrian RTI policies.

The majority of appraisals are carried out mid-term during one point in the programme’s lifetime. Mainly, a supportive purpose is followed as policy makers respectively programme managers need advice how to enhance programme implementation. Thus, appraisals focus mainly on policy/ strategy development, output counts and consistency matters. Whereas behavioural additionality issues are rather prominent in Austria, input and output additionality issues as well as quality of outputs are only considered in a limited number of evaluations. Technological, economic, and socio-economic impact dimensions are missing by large, or only refer to programme participants.

Looking at the data collection and data analysis methods employed, it seems that low cost data gathering methods (descriptive statistics and context analysis) prevail. Interviews and the usage of monitoring data are the most frequently used data collection methods in Austrian appraisals. Most commonly a mixed methodological approach where quantitative and qualitative methodologies are combined is used, although the number of methods is limited.

Compared with the other countries in the dataset, the most interesting difference refer to topics covered, impact dimensions covered, and data analysis methods employed. Only about one quarter of the Austrian appraisals cover input and output additionality issues whereas almost 50% of the appraisals in the other countries focus on these topics. Also the quality of outputs is widely neglected. Regarding impact variables, we witnessed that only a limited number of Austrian appraisals deal with impact at all. For every impact dimension coverage is lower in Austria than in the other countries of the dataset. If impact dimensions are covered, they rather focus on direct impact than on the participants than beyond.

The comparatively low coverage of impact dimensions, input and output additionality measurement, and consequently also econometric analysis might be due to the formative purpose of most evaluations. However, also case studies and context analyses are used significantly less in Austria than in other countries. Another reason for the discrepancies, which should not be neglected, is that the coverage of Austrian appraisals in the database is very high. Whereas in Austria almost the full range of appraisals conducted in the field of innovation policy is covered, it is more likely that only bigger evaluations are covered in the other countries; significant differences as regards the tender procedure (15% open tenders in Austria vs. 55% open tenders in the other countries) point in this direction.

Despite the intermingled picture as regards topics and impact dimensions covered and methods used, the quality of evaluations is perceived to be high by respondents. Given the evaluation purpose, also the methods used tend to be considered as appropriate. Especially, recommendations concerning changes to the management and implementation of RTI programmes were perceived to be useful. Forward-looking advice was regarded as helpful for the design and implementation of future policy measures.

To conclude with, we may state that in a period characterised by a catching-up process and structural change, Austria has become a country strongly positioned in the field of innovation policy and innovation policy evaluation. Evidence for the increased relevance of innovation policy evaluation was provided by changes in the legal conditions for evaluations, measures to foster an evaluation culture, the transparency of evaluation results, and the high number of evaluation activities. Austrian policy makers and agencies have used - or at least tried to use - evaluation as a tool to pursue a more rational policy approach. Nevertheless, due to the high number of evaluative activities, an increasing evaluation fatigue can be witnessed. Criticism was raised, that mechanisms ensuring that the results of evaluations do feed back into policy formulation and implementation are missing. In this respect, it will be important to spend more thoughts on the concrete purpose of planned evaluation activities and the role of evaluations for policy implementation, as it seems that too much of more of the same has been produced so far.

References

Aiginger, K., Falk, R. and Reinstaller, A. (2009): Reaching out to the Future Needs Radical Change. Towards a New Policy for Innovation, Science and Technology in Austria. Evaluation of Government Funding in RTDI from a Systems Perspective in Austria, Synthesis Report.

Arnold, E (2004) Evaluation of the Austrian Industrial ResearchPromotion Fund (FFF) and the Austrian Science Fund (FWF). Synthesis report. Wien.

CREST (2008), Expert group report on the design and implementation of national policy mixes, Policy Mix Peer Reviews: Country Report – Austria, http://www.bmwf.gv.at/fileadmin/user_upload/forschung/forschungsdialog/CREST_Austrian_Policy_Mix_Report_-_September_2008.pdf

Dinges, M. (2010) Öffentliche Forschungs- Technologie und Innovationsprojektfinanzierung, in Österreich: Ausmaß und Bedeutung im Innovationssystem, in Peter Biegelbauer (Hg.) (2010), Steuerung von Wissenschaft? Die Governance des österreichischen Innovationssystems. Studienverlag, Innsbruck/Bozen/Wien.

Edler, J. (2007), The Austrian Platform Research and Technology Policy Evaluation as a Forum of Strategic Intelligence. Views from abroad, in Platform Research and Technology Policy Evaluation and Austrian Council for Research and Technology Development (Eds.), Evaluation of Austrian Research and Technology Policies - A Summary of Austrian Evaluation Studies from 2003 to 2007, Vienna.

EIS, European Innovation Scoreboard (2008), Pro http://www.proinno-europe.eu/node/admin/uploaded_documents/EIS2008_Final_report-pv.pdf

Hirschhausen, C., Belitz, H., Clemens, M., Cullmann, A., Schmidt-Ehmcke, J., Zloczysti, P., (2009); Politikberatung kompakt 51, Innovationsindikator Deutschland 2009; Deutsches Institut für Wirtschaftsforschung (DIW), Berlin.

Pichler, R., Stampfer, M., Hofer R. (2007), Forschung, Geld und Politik: Die staatliche Forschungsförderung in Österreich 1945-2005, Studienverlag: Innsbruck.

Schibany, A. (2009), Der hinkende Frontrunner, InTeReg Working Paper Nr. 56-2009, Joanneum Research, Wien.

Schibany, A., Jörg, L., (2005), Instrumente der Technologieförderung und ihr Mix; InTeReg Research Report Nr. 37-2005.

Slipersæter, S., Lepori, B., Dinges, M. (2007) Between policy and science: research councils' responsiveness in Austria, Norway and Switzerland, Science and Public Policy, Volume 34, Number 6, July 2007 , 401-415(15).

Zinöcker, K. (2007), Evaluating Austria's R&D Policies. Some Personal Comments, in Platform Research and Technology Policy Evaluation and Austrian Council for Research and Technology Development (Eds.), Evaluation of Austrian Research and Technology Policies - A Summary of Austrian Evaluation Studies from 2003 to 2007, Vienna.

Zinöcker, K., Dinges, M. (2009), Evaluation von Forschungs- und Technologiepolitik in Österreich, in Widmer, Th., Beywl, W., Fabian, C. (Hg.), Evaluation – Ein systematisches Handbuch, VS Verlag für Sozialwissenschaften, Wiesbaden.

This page has been intentionally left blank.

Part III Chapter 9

Country Report: Germany

Stephanie Daimer and Susanne Buehrer[130]

---

[130] The lead authors for this section of the report are as given above, with contributions from all other team members at different parts of its production, conclusions and interpretations have been approved and are shared by the group.

Table of Contents

Table of Contents 266

List of Tables 267

Table of Exhibits 267

Executive Summary 268

1 Innovation policies and Evaluation practices at the federal level 269

2 Stocktaking of data collection and descriptive statistics 270

2.1 Policy Measure categorization 271

2.2 Commission of Evaluations 272

2.3 Types of Evaluations 272

2.4 Topics and impacts covered in the evaluations 273

2.5 Data collection and analysis methods 275

2.6 Audiences 276

2.7 Quality 277

2.8 Usefulness, Discussion, and Consequences 277

2.9 Linking Variables 278

3 An expert view on Evaluation practice in Germany 279

4 Conclusions: lessons to be drawn, shortcomings, and challenges 280

List of Tables

Table 97: German evaluated policy measures included into database 270
Table 98: Policy measure characterization: Modalities used in Germany compared to other countries 271
Table 99: Policy measure characterization: Target groups addressed in Germany compared to other countries 271
Table 100: Types of evaluations in Germany 2002-2007 272
Table 101: Timing and Purpose of German evaluations compared to rest 273
Table 102: Impacts covered in German evaluations 2002-2007 275
Table 103: Data collection methods in German evaluations (2002-2007) compared to the other countries 275
Table 104: Data analysis methods in German evaluations (2002-2007) compared to the other countries 276
Table 105: Quality of German evaluations (2002-2007) compared to the other countries 277

Table of Exhibits

Exhibit 80: Topics covered in German evaluations 2002-2007 274

Executive Summary

Four major findings for the innovation policies and the evaluation practice make Germany an interesting case to study and allow drawing some recommendations on good evaluation practice.

First, innovation policies in Germany are focussing on high technologies, SMEs and the still special situation of the Eastern federal states. This is clearly reflected in the evaluated policy measures in the Inno Appraisal database.

Second, the institutional setup at the federal level provides for quite a systematic approach to evaluation. Almost all programmes are being evaluated. In particular the Ministry of Economics regularly foresees evaluations, when planning new programmes. Open tender procedures and the commission of the evaluations to external evaluators are standard. This practice is not only clearly visible in the database. The INNO-Appraisal data show that this practice leads to particularly high quality of evaluations, specifically the application of open tender procedures is linked to high quality scores.

Third, evaluation reports are very often publicly available. There is particular interest in the evaluation community. The foundation of the Society of evaluation and several attempts of standardization have intensified scholarly debates. Actually, there is some sort of standardization of approaches visible, but more important, this convergent development takes place at a high quality level and includes the openness of evaluators (and commissioners of evaluations) towards new methods.

Finally, we have evidence from the data as well as from expert interviews that learning is a purpose of the commission of an evaluation. There are many formative evaluations, methods like focus groups or workshops are often employed, and the results of an evaluation are intensively discussed within government. Generally, it seems that learning actually takes place. But, although we find a high number of accompanying (and interim) evaluations in Germany, it seems that the learning applies in fewer cases to the evaluated measures themselves but takes place on a more general level namely the overall policy learning for future policy making and programme design. One of the reasons for this is that the aspect of "policy/strategy development" is an integral part of formative evaluations in Germany.

1 Innovation policies and Evaluation practices at the federal level

At least, since Benz and Daimler invented the car, Germany has constantly been the birth place of pioneering technological innovations, the latest of which is perhaps the mp3-format. According to the European Innovation Scoreboard (EIS), the innovation performance of the German economy is among the top within Europe and worldwide: It ranks the country as an "innovation leader" at the 3rd position within the EU in 2008.[131]

Science, Technology and Innovation (STI) policies are central topics for the policy makers. This and the fact, that competences for these and related policy areas such as education or economic policies are shared by the federal and regional levels leads to an enormous number of political activities and measures. However, still the R&D expenditures remain below the EU's goal of 3% of the GDP and it is not very likely that it will be reached by 2010.

The challenges at hand and the main political priorities for the federal government include a focus on SMEs, on regional clusters and on the still particular situation of the Eastern Laender. Moreover, as regards the education system and the rise of new technologies, Germany is internationally less competitive, which is why these topics are high on the agenda and new mega-measures like the high-tech-strategy have been created. The "High-Tech-Strategy" may be regarded as the current de-facto national innovation strategy. Since February 2008, there is also now a "Strategy for Internationalizing Science and Research". The Christian-democratic-liberal coalition government, which has come into office in autumn 2009, has not yet announced any plans for major changes to the innovation policies of the grand coalition government.

There are two main actors in STI policies at the federal level: the Federal Ministry of Economics and Technology (BMWi) and the Federal Ministry of Education and Research (BMBF). They sponsor the innovation policy measures, while there is no co-sponsorship by EU structural funds at the federal level. The programme administration itself is regularly handled by external programme managers: several public or private agencies deal with programme implementation and administration while the ministries define the policy objectives and focus on the strategic design of the measures.

Innovation policy and the evaluation of the policy measures are strongly intertwined. Evaluations are often foreseen, already when designing policy measures. In particular, all BMWi programmes undergo interim and / or ex-post evaluations, and sometimes there is additional accompanying research. The BMBF approach has been less systematic; there are a number of thematic R&D programmes which run without being (externally) evaluated. In the meantime, also the BMBF regularly plans external evaluations. So nowadays, open tender procedures and the commission of an evaluation study to an external agency or institute are standard. The appraisal reports are often publicly available, however often only in German.

There is a vivid scholarly interest in the evaluation of public policy. The Society for evaluation (DeGEval) has been founded in 1997 and German researchers are active in the international scientific community for evaluation research. Evaluation studies often refer to scholarly defined common standards or guidelines defined by the programme owners, which is why one can observe some sort of convergence regarding the design of evaluation studies (e. g. coverage of certain topics or impacts, methods of data collection or analysis …).

---

[131] European Commission, DG Enterprise and Industry. 2009. European innovation scoreboard 2008 Comparative analysis of innovation performance. January 2009. http://www.proinno-europe.eu/www.proinno-europe.eu/admin/uploaded_documents/EIS2008_Final_report-pv.pdf.

2 Stocktaking of data collection and descriptive statistics

For Germany, a sample of 18 evaluations has been identified. The German cases make up the second largest country group in the dataset, following the large set of Austrian evaluations. These 18 evaluations are covering 14 different policy measures of the federal government. All 18 prefilled templates have been amended and validated by policy makers. For the German case, the Trendchart database represents a valid source. Additional desk research has shown that all evaluations relevant for the Inno Appraisal database are mentioned in the PRO INNO Trendchart database of policy measures. Table 97 gives an overview of the policy measures included into the Inno Appraisal database. Three policy measures are covered by more than one evaluation. All the evaluations considered here are single measure evaluations, so the German sample does not include any portfolio or structural fund evaluation. Eight programmes fall into the responsibility of the Research Ministry (BMBF), among them mainly sector-specific R&D programmes and two programmes aiming at regional growth. The six programmes owned by the Economic Ministry (BMWi) include SME-funding and different forms of science-industry cooperation as well as start-up funding.

Table 97: German evaluated policy measures included into database

Policy measure	TC category	Evaluations	Publication	Ministry responsible
InnoRegio - innovative networks in Eastern Germany -incl. Interregional Alliances-	DE16	1	2005	BMBF
INNO WATT - Special R&D programme for Eastern Germany	DE19	1	2006	BMWi
EXIST - Start-ups from Science	DE21	1	2006	BMWi
InnoNet	DE26	1	2004	BMWi
PRO INNO II	DE28	3	2005, 2006 (2)	BMWi
INNOMAN: Innovation management in SMEs in Eastern Germany	DE33	1	2007	BMWi
FUTUR -Technology Foresight, former- Delphi 1998-	DE35	1	2005	BMBF
Applied Research at Universities of Applied Sciences in Co-operation with Business (FH3)	DE42	1	2008	BMBF
Innovative Regional Growth Poles	DE57	1	2005	BMBF
Thematic R&D programs - Production technology	DE68	2	2003, 2006	BMBF
Thematic R&D programs - Plasma technology	DE68	1	2004	BMBF
Thematic R&D programs - Mikrosystems technology	DE68	1	2003	BMBF
Thematic R&D programs - Bioregio & Bioprofile	DE70	1	2007	BMBF
NEMO - Management of Innovation Networks for East German SMEs	DE75	2	2005, 2007	BMWi

Legend: TC = PRO INNO Trendchart; Publication = Publication date of the evaluation report(s), BMWi=Federal Ministry for Economics, BMBF=Federal Ministry for Education and Research.

2.1 Policy Measure categorization

The Inno Appraisal database characterizes policy measures according to their modalities and target groups. This characterization is based on categories employed in TrendChart. We have reduced the number of modalities in order to minimize multiple response coding to a maximum of three modalities. Table 98 displays the characterization of the German policy measures. Most of the measures combine two or three modalities. The most dominant type of measures are so-called multi-actor programmes (see modalities M7 and M8), which support the cooperation of Universities or public research organizations with companies. Nine measures either focus on the support of networks and clusters (M7) or on science-industry cooperation in general (M8). Most of these measures provide support by direct funding (M2). Non R&D related support (M3) is mainly directed towards innovation management. This category also includes one Technology Foresight programme (FUTUR). The table shows that Germany has no tax measure scheme (M1), although this is hotly debated at the moment. It also shows that there the sample does not include a special start-up support measures such as an incubator (M6). Start-up support is granted indirectly via intermediary agencies (M4) located at universities and funded by the EXIST programme. Two measures intend to promote the mobility of personnel (M5).[132]

Table 98: Policy measure characterization: Modalities used in Germany compared to other countries[133]

Modality of policy measure	Frequency in DE	Frequency in other countries
Indirect measures (M1)	0%	5%
Direct financial support (M2)	72%	57%
Non-R&D related support (M3)	33%	30%
Creation of intermediary bodies (M4)	17%	6%
Mobility of Personnel (M5)	22%	7%
Creation of start-ups (M6)	0%	10%
Networks & Clusters (M7)	39%	29%
Science-industry cooperation (M8)	22%	25%
Support for the uptake and diffusion of innovation (M9)	6%	24%

Table 99: Policy measure characterization: Target groups addressed in Germany compared to other countries[134]

Target Group of policy measure	Frequency in DE	Frequency in other countries
Universities, Public Research Organizations	83%	66%
All firms	44%	60%
SMEs only	50%	19%
Sectors	39%	17%
Regions	44%	11%

---

[132] Significant relationship is mainly due to the fact, that in total 4 out of 18 evaluation reports refer to this category, because there are three evaluation reports for the policy measure PRO INNO II.

[133] Shaded cells show significant associations between the column and row variables.

[134] Shaded cells show significant associations between the column and row variables.

Compared to the other countries in the sample there are three specifics in the German innovation policy with respect to the addressed target groups (Table 99). At a statistically significant level, Germany has more measures targeted at SMEs, at certain sectors and at special regions. This finding goes in line with the identified challenges of the innovation system as described above. Special focus has to and is being paid to the prosperity of SMEs, to high-tech sectors, to regional clusters as well as to the whole region of Eastern Germany.

2.2 Commission of Evaluations

The Inno Appraisal database mirrors the general knowledge about evaluation practices in Germany. All evaluations in the database are done by external evaluators. Additionally, we find a high number of open tender procedures, 15 in total. Both results are statistically significant results compared to commission practices in other countries. 13 Evaluations were already foreseen, when the programme was designed, which is due to the rules established by the Ministry of Economics. For all five evaluations which were commissioned without prior planning, the Research Ministry was responsible for. For six evaluations, there was in addition a pre-defined budget.

2.3 Types of Evaluations

Among the variables the Inno Appraisal database collects for each appraisal, the aspects of timing and purpose classify an evaluation study rather good. The project distinguishes among four types of timing, ex-ante, ex-post, interim and accompanying. Accompanying evaluations takes place at several points in time, while an interim evaluation takes place at a single moment during the runtime of a program (e.g. at the end of a budget period). Ex-post is defined as taking place after the termination of the evaluated measure. In most of the analysed appraisals there is a blend of formative and summative purposes, however, one element is often dominant and therefore chosen, when coding.

Evaluations of German innovation policy measures so far seem to be either accompanying, interim or ex-post, but there are no ex-ante evaluations in the sample. Moreover the cross-tabulation in table 100 shows, that the combination of timing and purpose clearly reveals three types of evaluations in Germany. The ex-post evaluations follow mainly a summative purpose. Secondly, a larger group of studies are accompanying evaluations. When accompanying evaluations are commissioned, the ministries often intend a formative approach in order to be able to redesign certain aspects of ongoing policy measures. The third group are interim evaluations of long-term measures with a primarily formative character. This makes in total 72% of formative evaluations, which is quite a high share.

Table 100: Types of evaluations in Germany 2002-2007

Timing Purpose	Accompanying	Interim	Ex-post	Total
Formative	33%	39%	0%	72%
Summative	0%	0%	28%	28%

The differences between Germany and the other countries in the sample for the timing and purpose of evaluations are statistically significant (see Table 101). Based on this result one might get the impression that German policy-makers seem to be interested in policy improvement through intelligent information delivered by evaluations. Additionally, the political and scientific discourse in the German evaluation community supported the prevailing paradigm of "learning" and formative evaluation approaches too.

Table 101: Timing and Purpose of German evaluations compared to rest[135]

		Frequency in DE	Frequency in other countries
Timing	**Ex-ante**	0%	14%
	Accompanying	33%	12%
	Interim	39%	43%
	Ex-post	28%	28%
Purpose	**Summative**	28%	20%
	Formative	72%	38%
	Both	0%	37%

2.4 Topics and impacts covered in the evaluations

Exhibit 80 gives an overview on the topics the evaluations cover. This picture mirrors by and large the picture of the whole sample. Most of these topics are not employed more or less often in Germany compared to other countries at a statistically significant level.[136]

All the evaluation studies cover output or outcome issues and evaluate whether the measures' goals have been achieved. Often, the quality of outputs is assessed (e.g. patents, prizes, excellence rankings).

Many evaluations cover context analysis topics, such as coherence/complementarity with other policy measures or the external consistency of the measure. The coverage of efficiency issues, in particular programme implementation efficiency is also quite frequent. Also, more than half of the studies aim at contributing to policy and strategy development.

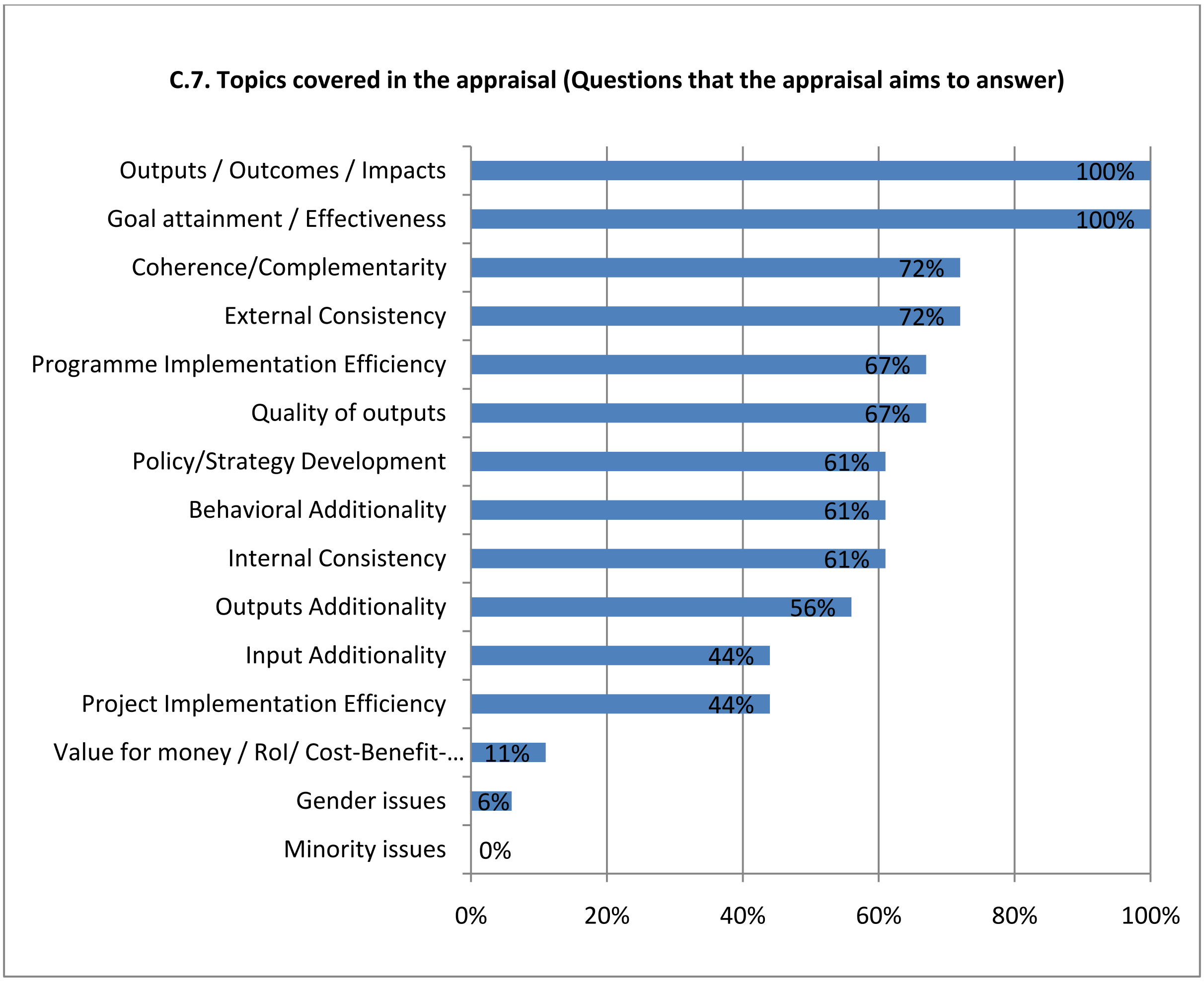
Among the additionality topics, behavioural additionality ranges rather high. This result can be attributed to the fact that formative approaches are the most frequent ones in Germany. These are typically conducted rather early in the lifecycle of a programme when concrete outputs are hardly to identify. A further reason can be seen in the dominant type of (network and cluster oriented approaches) where we also expect - in the first phase - primarily a behavioural change and not so much a concrete output in terms of new products or processes.

---

[135] Shaded cells show significant associations between the column and row variables.

[136] Internal consistency and Gender issues are exceptions. Both topics are less often employed in Germany compared to the rest of the sample.

Exhibit 80: Topics covered in German evaluations 2002-2007



Looking at the impact variables in the database, we see, that a lot of studies in Germany cover impact dimensions in a broad perspective, which means "on the participants and beyond" (see Table 102). Looking at both levels of aggregation, economic and technological impacts are considered in almost all evaluations. Social, or rather societal impacts are also of high interest in Germany (in almost every second case). In most cases, the creation of new jobs is being covered under this topic. Scientific impacts' coverage could have been expected to be covered more often, as there are many programmes targeted towards science, too. But – as we saw above – most programmes in the German sample are designed to improve the exchange between the science sector and companies and consequently we observe effects in the cooperation behaviour and technological development which then is expected to lead to improved economic impacts. The technology focus in Germany is by far stronger compared to the rest of the sample, where 56% of the evaluations cover technological impacts. Environmental impacts seem to be of a lower interest (so far) in Germany. In other countries, this impact type is covered on average in every third evaluation.

Table 102: Impacts covered in German evaluations 2002-2007

Level of Aggregation	On the participants (direct)	On the participants and beyond (indirect)	Total
Economic	17%	67%	84%
Technological	22%	56%	78%
Societal	0%	44%	44%
Scientific	6%	28%	34%
Environmental	0%	6%	6%

2.5 Data collection and analysis methods

There seems to be a standard for the data collection of evaluations, combining quantitative and qualitative data (see table 103). These include the usage of monitoring data provided by the programme managers as well as participant surveys and additional interviews (often with experts) by the evaluators themselves. Internationally compared, these three data collection methods are also quite frequent, however on average less frequently used than in Germany. In other countries, existing surveys and databases and document search are more often used than in Germany, where they appear in about half of the cases.

Workshops are a method clearly associated to formative evaluations and many German evaluations employ this method. This is clearly a difference to the rest of the countries in the sample.

Non-participant surveys, peer reviews and technometrics/bibliomentrics research clearly appear to be rarely considered methods in Germany as well in international comparison. Even although quality of outputs is often assessed in Germany, none of the studies applies rigorously technometrics/ bibliometrics research. One reason could be that due to technical particularities it takes a certain amount of time until bibliometric or other effects are measurable within the relevant databases (like the European Patent Office etc.) and this does not fit with formative/interim evaluations.

Table 103: Data collection methods in German evaluations (2002-2007) compared to the other countries[137]

	Frequency in DE	Frequency in other countries
Monitoring Data	100%	77%
Participant Survey	83%	63%
Interviews	83%	75%
Focus Groups / Workshops	72%	47%
Existing Surveys / databases	50%	72%
Documents	44%	67%
Peer Reviews	33%	17%
Non-Participant Survey	22%	25%
Technometrics / Bibliometrics	0%	2%

Looking at table 104, data presentation in German evaluations regularly includes descriptive statistics and context analysis, as it is also the case internationally. However, the application of these methods is quasi-systematic in Germany, which makes a difference compared to the rest of the sample.

---

[137] Shaded cells show significant associations between the column and row variables.

Apart from that, data analysis methods are quite different in Germany compared to the rest of the countries. For qualitative analyses, case studies are used in Germany as often as internationally, but German evaluations apply remarkably less document analysis – understood in a strict sense as the systematic analysis of variables which have been coded using a defined set of documents. Quantitative approaches play also a role in Germany, although we find here often econometric analyses, while in other countries more specific methods such as input/output analyses are more important. Among the more experimental designs, control group and counter-factual approaches are rather frequent. There is no tradition for Cost-Benefit Approaches in Germany.

Network analysis is quite seldom, which is somewhat surprising, as there are many network policy measures in the sample. But this finding is not different from the whole sample. One reason might be that a real network analysis requires a large amount of resources (time, money) and a clearly defined set of actors which is not always given in the respective programmes.

Table 104: Data analysis methods in German evaluations (2002-2007) compared to the other countries[138]

	Frequency in DE	Frequency in other countries
Descriptive Statistics	100%	73%
Context Analysis	89%	64%
Case study Analysis	39%	42%
Econometric Analysis	39%	21%
Counter-Factual Approach	33%	20%
Control Group Approach	28%	19%
Document Analysis	17%	56%
Network Analysis	11%	18%
Input/Output Analysis	11%	28%
Quasi-Experimental Design	6%	11%
Cost-Benefit Approach	0%	26%

2.6 Audiences

In general, the main audiences of the evaluations in Germany are like in the whole sample, policy makers and programme managers, and to some extent also politicians and parliamentarians (auditors). We find three differences compared to the rest of the sample. External co-sponsors are not addressed by German evaluations, which is explained by the fact that there is no external co-sponsoring. More surprisingly, policy analysts are very often addressed by evaluations, a statistically significant difference to other countries. This coincides with the description of the German evaluation practice that includes a vivid scholarly interest and debate about evaluation methods and research, and a professional Society (DeGEval) that offers the respective fora for exchange. Thirdly, the general public is a main audience for many evaluations in Germany as well as are those directly supported by the measure and potential users of the measure. Evaluation reports seem to be important for external communication, too.

---

[138] Shaded cells show significant associations between the column and row variables.

2.7 Quality

Quality assessments have been made by policy makers. According to them, the overall quality of the appraisals in Germany is at a very high standard. The highest quality scores are related to the issues "evaluation addresses the terms of reference (TOR)", "analysis based on the given data" and "conclusions based on analysis".

The studies often rely largely on descriptive statistics. There is regularly a clear documentation of data sources included. Finally, context coverage is quite good, which corresponds to the frequent application of context analysis and the broad coverage of impacts. However, if other methods, like econometrics, are used, their documentation is sometimes rather brief.

Compared to the other countries in the sample, quality scores for German evaluations are on average higher for all aspects of quality. In many cases, this difference is statistically significant (see Table 105).

Table 105: Quality of German evaluations (2002-2007) compared to the other countries[139]

	Mean in DE	**Mean in other countries**
Address TOR	4,83	4,10
Design appropriate given the objectives	4,39	4,05
Methods satisfy the TOR/purpose	4,33	4,32
Application of qualitative methods	4,12	3,94
Application of quantitative methods	4,06	3,83
Information sources well documented	4,56	4,07
Analysis based on given data	4,83	4,20
Cover broader context	4,12	3,40
Conclusions based on analysis	4,89	4,26

2.8 Usefulness, Discussion, and Consequences

Compared to the rest of the database, German evaluations show some differences in the perceived usefulness of the recommendations. The template differentiated between the following categories:

Internal usefulness

- Changes to the design of the programme/measure appraised
- Changes to the management and implementation of the programme/measure appraised

External usefulness.

- Changes to the design, management and implementation of future programmes/measures
- Changes to the design, management and implementation of contemporaneous programmes/measures
- Changes to broader policy formulation and implementation

---

[139] Shaded cells show significant results of T-test.

On average, the external usefulness is higher (which is true for the design of future programme and broader policy formulation, but does not include the design of contemporaneous programmes), and the results are mixed for internal usefulness (higher for design of measure, lower for implementation issues). Statistically significant are the differences for two aspects of external usefulness: future measures and policy formulation. To find higher scores for usefulness fits the above finding on the number of formative evaluations in the database.

To sum up: Generally, it seems that learning is a purpose of the commission of an evaluation, and takes actually place. But, although we find a high number of accompanying (and interim) evaluations in Germany, it seems that the learning applies in fewer cases to the evaluated measures themselves but takes place on a more general level namely the overall policy learning for future policy making and programme design.

German evaluations are discussed on average more within government and slightly more with stakeholders compared to other countries. The level of discussion within government is to a statistically significant extent higher than in other countries. This again confirms the finding of the learning aspect.

When it comes to measurable consequences, we find that the expansion and prolongation of a measure follows in Germany quite often from evaluations. Although this is a frequent consequence for the whole database, the result for Germany is significantly higher. This might be due to the fact that the public R&D expanses at the federal level increased from 9 billion in 2001 up to more then 11 billion in 2008.

2.9 Linking Variables

In this section, we are looking at relationships between the variables. The results will give some more illustration of the evaluation practice in Germany, although they have to be interpreted with care, as they are based only on 18 observations.

For the three identified types of German evaluations we find some linkage to certain topics and methods. In particular we find statistically significant relationships for formative evaluations with the topics “External Consistency”, “Coherence/Complementarity”, "Quality of outputs", "Programme implementation efficiency" and “Policy/Strategy development” and with the methods "participant surveys", "interviews" and “focus groups/workshops”. These evaluations tend to apply classical methods, while summative seem to employ rather different methods. As a main approach, the counter-factual approach and control group approaches are linked to summative evaluations as well as input additionality, what might allow the conclusion that they put more weight on impact assessment, while formative also cover other topics such as implementation of the measure. Related to impact assessment, we find that studies which evaluate direct financial support measures (Modality 2) cover economic impact assessment, often including indirect economic impacts.

We have also tested the relationships of characteristics of the evaluations with the perceived quality and usefulness by the policy makers. There are three results with statistical significance:

- One result is that open tender procedures are linked to top quality scores for the aspect "conclusions are based on the analysis". This indicates that open tender procedures can contribute to the quality of an evaluation.

- A second finding shows that the coverage of technological impacts is linked to higher quality scores for the "coverage of the broader context". In our view this can be interpreted as a hint that evaluations of technology programmes in Germany take into account the system perspective and actor constellations to a great extent
- Finally, we find that the topic of "Policy/ Strategy development" is linked to the usefulness for the "design/ management/ implementation of future programmes". This again points to the fact that German evaluations have a high external usefulness, in particular those which aim to explicitly contribute to policy / strategy development.

3 An expert view on Evaluation practice in Germany

We can interpret and qualify further our findings from the INNO-Appraisal database using results from in-depth interviews with German experts. We interviewed five policy makers or programme officers from ministries and project agencies, who represent a broad experience with commissioning evaluations, acting as an evaluator themselves and implementing evaluation results.

We asked the experts to reflect about the evaluation practice and whether they have been observing changes during the last couple of years. The experts overall agree that the evaluation practice is at a high level and one might talk of an evaluation culture in STI policies. All experts have observed changes during the last couple of years with a clear trend towards more professionalization and standardization.

On the side of the policy makers this means that they in general know very well what to expect from evaluation studies and that they are well aware of methodological limitations for example for impact assessments. There is a clear trend towards accompanying and interim evaluations, which indicates according to the experts that evaluation is being used as a learning tool.

On the side of the evaluators, this means that a community has developed which performs evaluations at a high quality level. There is interest and participation in (international) scholarly debates, and an interest in new methods such as evaluation econometrics, which leads to convergence and even standardization in evaluation concepts and methods applied.

The results from the database reflect the view of the experts. We found for Germany a very systematic approach to evaluation including open tender procedures and external evaluations. During the observation period, this has been true for the evaluation of innovation policy programmes owned by the Economic ministry (BMWi), but in the meantime also the Research Ministry (BMBF) has become more systematic and plans evaluations for all programmes. The impression of the interviewed experts regarding quality confirms what has been found in the data: evaluations in Germany are of a remarkably high quality and one of the reasons for this is the practice of open tenders.

We see in the data a dominance of accompanying research and a clear linkage of this evaluation type with a formative evaluation approach, which underlines the importance of learning from evaluation. We have also seen that evaluation studies in Germany very often address policy analysts, which underlines that the evaluation community is interested in scholarly debates.

4 Conclusions: lessons to be drawn, shortcomings, and challenges

The evaluation practice of innovation policy measures in Germany is at a high level. All programmes are being evaluated, however, some of them only internally.

While the transparency for external evaluations is high, meaning that reports are in general available, it is not possible to trace the internal evaluations. If mentioned at all, this can be found in Trendchart, however, without any available information about the results. Documentation of internal evaluations via programme owners/managers is also not accessible. This leads to a quite homogeneous sample for Germany, resulting in a set of external evaluation studies, which have been commissioned in open tenders.

Obtaining policy maker responses has brought mixed results. For some programmes, reallocation of responsibilities has caused discontinuities: the responsible programme officers have changed between the two ministries, BMWi and BMBF, or within the ministries. Policy makers who do not know the programmes for their whole life-time are more reluctant to answer our request. On the other hand, quick reactions and detailed responses of policy makers show that the project appeals to the policy makers and is also an indicator that the validation of pre-filled templates by policy makers can bring about the intended effect of high data quality.

The most interesting differences of German evaluations compared to the rest of the sample concern first the target groups and the types of evaluations. We find in particular many programmes targeted towards SMEs, sectors and regions. This finding goes in line with the identified challenges of the innovation system. Special focus has to and is being paid to the prosperity of SMEs, to high-tech sectors, to regional clusters as well as to the whole region of Eastern Germany. Moreover, the combination of timing and purpose clearly reveals three types of evaluations in Germany: formative accompanying, formative interim and summative ex-post studies. To a certain extent, also certain topics and methods are associated with these types. For example, classical data sources such as participant surveys and interviews are mainly used in formative evaluations. A classical formative data source – focus groups and workshops – are also used very often in Germany. Summative evaluations on the other hand are linked to counter-factual approaches, a classical impact assessment approach.

Apart from that, impacts are being covered very often at an indirect level (beyond the participants). Looking closer at the impacts covered, we do not find yet, that new impact types such as gender aspects, sustainability aspects, public health, innovation mentality or risk attitudes are widely spread. However, from in-depth expert interviews we know, that the evaluation community has quite a critical view on the mere focus on economic aspects: There is neither often the possibility to assess them, because of problems such as the time-lag of effects, nor does the economic dimension fully cover today’s objectives associated with innovation policy. Today, we witness the trend that innovations do have to bring about more or other effects than economic success and growth: (normative) societal needs are becoming more important (see for this also the chapter on impact assessment). Policy makers will have to pay even closer attention in the future, what kind of objectives they wish to pursue with a certain measure. Evaluators on the other hand will be challenged with new impact types and the question how to measure them.

In terms of quality, the German evaluations range at a very high level – significantly above the average of the sample. German policy makers obviously value the work of the evaluators. This is supported by findings for usefulness and consequences. German evaluations contribute in particular with respect to aspects of external usefulness (e.g. design of future policy measures and broader policy formulation). Quite frequent, the expansion and prolongation of programmes is a consequence of evaluation reports.

To conclude in more general terms, we find from the INNO-Appraisal database and from interviews with experts that Germany has quite a developed, professional and standardized approach to the evaluation of innovation policy measures. There is a general trend towards more accompanying and formative research, which underlines the fact that evaluation is being used as a learning tool for policy makers.

This page has been intentionally left blank.

Part III
Chapter 10

Evaluation in the United Kingdom

The UK is widely held to offer a good example of a governance system with a strong evaluation culture. This Chapter examines the particular context within which the evaluation of innovation policy support resides in the UK system of governance and how these have developed both historically and in response to broader policy concerns.

It then presents the main features of the current processes, tools and structures that frame evaluation in the UK.

Finally, it examines a number of specific issues, such as the selection of evaluators, dedicated budgets and planning for evaluation, the use of recommendations, etc, that were investigated by the INNO Appraisal survey of evaluation reports and offers examples of how these are approached in the UK context.

Paul Cunningham and John Rigby[140]

---

[140] The lead authors for this section of the report are as given above, with contributions from all other team members at different parts of its production, conclusions and interpretations have been approved and are shared by the group.

Table of Contents

Table of Contents 284

Table of Exhibits 285

Executive Summary 286

1 Introduction 288

1.1 Strategic Review 288

1.2 Multiple actors 291

1.3 The shift in innovation policy support 291

2 A strong evaluation culture 292

3 The historical context shaping evaluation 293

3.1 Systematic Approaches 294

3.2 Learning by Doing – Towards Meta-evaluation 294

3.3 Consensus about Economic and Social Models 295

3.4 Broadening Evaluation and Appraisal in Government 295

4 Current practice 296

4.1 Overview 296

4.2 The Green Book 297

4.3 Business cases, programme plans and balanced scorecards 299

4.4 The Magenta Book 302

5 Specific issues from the survey 303

5.1 Rationale and purpose 303

5.2 Evaluators 303

5.3 Terms of Reference and the use of innovative approaches 303

5.4 Timing 304

5.5 Conditionality 304

5.6 Dedicated budget for evaluation 304

5.7 Is evaluation foreseen and planned? 304

5.8 Topics, data collection methods, and data analysis methods 304

5.9 Programme impacts 305

5.10 Sponsors, audiences and availability of results 305

5.11 Recommendations 305

5.12 Quality and utility 305

6 Conclusion 306

Table of Exhibits

Exhibit 81: The ROAME-F Cycle 297
Exhibit 82: Comparison of Project Life Cycle Stages 298

Executive Summary

It is a widely accepted belief, supported by documented evidence, that the UK has a strong culture of evaluation in RTDI policy making. This case study examines the broader context within which the processes of review, assessment, appraisal, monitoring and evaluation are employed within the UK system of innovation policy governance, a system which, due to the broad definition of innovation held in the UK, encompasses a number of policy domains and actors.

In particular, a number of relevant features of the UK innovation policy governance system are considered, including:

- The use of strategic review processes (and a framework for performance monitoring)
- The presence of multiple actors and stakeholders
- Multi-level governance
- The evolutionary shift from direct support to framework support.

The study then looks at the underlying factors and developments that have shaped the evolution of the current system of evaluation practice in innovation policy governance. These are: a) the development of systematic approach to evaluation in the 1970s and 1980s; b) the accumulation of evaluation expertise through limited meta-evaluation that has led to an innovation culture in government which recognises the value of a practical business oriented approach to policy; c) the growing consensus around the neoclassical model of the economy and society; and d) the extension of evaluation activities throughout government as the devolution of policy and programme and project design and their evaluation has been pushed downwards and outwards from Whitehall to the regions.

Current evaluation practices and tools are then reviewed, in the context of recent structural changes in the machinery of governance in the UK, with a focus on those employed by the Department for Trade and Industry (DTI) and its more recent incarnations, the Department for Innovation, Universities and Skills (DIUS) and now the Department for Business, Innovation and Skills (BIS). The overarching influence of HM Treasury across all policy domains (and the imperative of demonstrating ‘value for money’ from policy intervention) is exemplified by the guiding principles set out in its ‘Green Book’, whilst the promotion of a systematic approach to the policy cycle and to performance measurement (including the use of appraisal, monitoring and evaluation) is underlined by the use of tools such as business cases, programme plans, balanced scorecards and the ROAME-F tool. Evidence is also provided for the cascading down of this guidance to the regional level of governance.

There is also support for the fact that policy interest in the UK extends beyond the mundane and routine application of evaluation as a formalised requirement and into the more exploratory and learning-oriented application of evaluation as an evolving policy tool which is adaptable to a variety of new and changing contexts. This is evinced by the ‘Magenta Book’, which provides guidance on social research methods for policy evaluation and endeavours to develop a greater understanding of the use and applicability of various approaches to evaluation, from the broad to the specific level.

Overall, it is clear that there is an extensive literature and a range of embedded practices relating to appraisal and evaluation in the UK policy system, all of which reinforces the view that the country possesses a well developed evaluation culture.

The study ends with a more detailed examination, in the UK context, of a number of issues which the INNO Appraisal survey of evaluation reports sought to investigate. These were:

- The rationale and purpose for an evaluation: primarily this is aimed at ensuring value for money, coupled with policy learning considerations, which can include identifying unanticipated outcomes and spill-over effects.
- The sourcing and selection of evaluators: all evaluators are external, ensuring independence and evaluation competence, with open tendering a preferred option. Evaluators must meet stringent criteria.
- The use of terms of reference and opportunities for innovative evaluation approaches: Terms of reference are set according to established principles; exploratory approaches are encouraged, provided the principal requirements for the evaluation are met.
- The timing of evaluations: depends on context – the rolling nature of UK programmes tends to favour interim evaluation. Monitoring and appraisal are also standard practices.
- The conditionality of evaluations: evaluation is a pre-condition of HM treasury funding for interventions above a certain funding level.
- The use of dedicated budgets for evaluation: Evaluations are always foreseen and budgeted for.
- Planning of evaluations: All programme formulation includes appraisal, monitoring and evaluation as anticipated elements.
- Topics, data collection methods and data analysis methods: these are all highly dependent upon the context and purpose of the innovation support measure under evaluation. The Magenta Book offers guidance on the appropriate methodologies for use.
- Programme impacts: Evaluations tend to look for both anticipated and unanticipated impacts. Again, the Magenta Book provides guidance on programme impact and how it may be measured.
- Sponsors, audiences and the availability of results: Programme managers form the immediate audience although HM Treasury is the ultimate audience and sponsor. Evaluation in BIS is also under scrutiny from a high level steering group. As a rule, all evaluation reports are made publicly available, except in certain cases where confidentiality concerns arise.
- The production and uptake of recommendations: Recommendations, provided they are realistic and economically feasible are generally acted upon. Similarly, they will be published provided confidentiality concerns do not arise.
- Quality and utility: Quality is defined as being fit for purpose, meeting the Terms of Reference and delivering within budget. Quality is an asymptotic function: there is a minimum level of quality that must be achieved for the delivery of the evaluation’s objectives. An evaluation is deemed to be useful if the evaluation delivers the Terms of Reference in a consistent manner and if it provides actionable recommendations and delivers value for money

In conclusion, it is clear that the UK does indeed possess an extensive and historically well-developed culture of evaluation which though formalised and set firmly in a framework geared towards the assessment of performance measurement, policy relevance and value for money, is nonetheless adaptable, context sensitive and reflexive and, moreover, practised by a policy community that appreciates it as a key tool for policy learning and improvement.

1 Introduction

The UK is frequently considered to be a country which demonstrates a strong ‘evaluation culture’. Indeed, the use of evaluation (and the accompanying practices of assessment, appraisal and monitoring) forms part of a well-established focus on the use of a range of governance processes in policy-making, especially in the areas of R&D and innovation support. For example, a recent CREST OMC 3% Policy Mix Peer[141] review of the country highlighted the UK for its use of review and stakeholder engagement, coupled with a well developed culture of evaluation, in the formulation of policies. Another feature of this emphasis on governance is that, for several years, the UK has made its policy transparent through the publication of long-term strategic documents backed up with clear implementation plans and regular monitoring of progress.

It is also worth noting that an important general feature of the UK’s R&D policy-making processes concerns the broad definition of innovation that is employed within government. This definition encompasses not only technology-based innovation, but also innovation in terms of management practices, service provision, business models, etc. Such a definition covers many of the innovations that occur in the service sector and does not restrict itself to the more usual technology-based innovations found in the manufacturing sector. Moreover, S&T and R&D policy concerns are embedded within innovation policy, thus a broad range of stakeholder views, from both within Government and outside it (from the private, public and not-for-profit sectors) are taken into consideration during the formulation of innovation policy.

It is therefore important to consider these features when discussing the evaluation of innovation policy support, as such mechanisms exhibit a wide range of targets, objectives and modalities.

In addition, a number of features of the UK innovation policy governance system are also relevant when considering the process of evaluation. These include:

- The use of strategic review processes
- The presence of multiple actors and stakeholders
- Multi-level governance
- The evolutionary shift from direct support to framework support

A brief account of each of these features is described below.

1.1 Strategic Review

The publication of the UK’s Ten-Year Science and Innovation Investment Framework (SIIF) 2004-2014 in 2004 marked the latest culmination of a series of in-depth reviews of the UK’s system of innovation (see below). The strategy outlines the UK Government’s planned investments into S&T and innovation policy-related activities over a ten year period.

Of particular note, was the typical use of the process of **in-depth system-wide review** undertaken in drawing up the Framework, in which the Government consulted extensively with key stakeholders. These included the scientific community, businesses, charities and regional and devolved bodies, as well as international contacts. Around 200 contributions were received from a wide range of individuals and organisations. In order to set the agenda for these contributions, a consultation document *Science & innovation: working towards a ten-year investment framework* was released with the 2004 Budget in March 2004. This drew on a series of existing reviews and analyses across the existing policy mix.

---

[141] Cunningham, P.N. CREST 3% OMC Science & Innovation Policy Mix Peer Review: United Kingdom, Background Report, European Parliament, March 2007.

A second feature was the **engagement of a wide range of Government stakeholders** (which also had substantial public sector R&D funding and policy making responsibilities) in the delivery of the Framework. These included the Treasury (responsible for all Government spending), the Department of Trade and Industry (responsible for the Science Budget and several innovation support measures), the Department for Education and Skills (responsible for university block funding), the Higher Education Funding Councils, the Research Councils and the then Office of Science and Technology, together with other Government departments with significant scientific and technological portfolios. The Framework was also jointly published by the Treasury, the DTI and the DfES, thereby underlining the coordinated approach to policy strategy, although the DTI (already recognised as the key agency for innovation policy matters) was identified as the lead agency for taking the Framework forward and for monitoring progress. The new Department for Business, Innovation and Skills (BIS) is taking forward this role.

Of particular relevance to the broader context within which evaluation takes place is the underlying performance monitoring system of **Public Service Agreements** (PSAs). Put in place by the Treasury (the UK's ministry of finance), this serves as a broader mechanism for performance measurement and for monitoring progress against targets. Held by all responsible agencies (i.e. ministries), failure to meet PSAs can affect future budgetary allocations (allocated through three-year Spending Reviews – see below); hence it is in the clear interest of ministry officials to ensure that their policies are designed to effectively and efficiently meet Treasury targets. A range of stakeholders may be consulted on the technical and operational details of policy measures, depending on the type of measure being designed. For example, fiscal measures will involve major inputs from HM Treasury and the Inland Revenue, while technology transfer measures will take account of the views of business representatives, universities, intermediary organisations, employers' representatives, etc. The way in which this involvement is handled will vary on a case by case basis.

PSAs serve as clear targets which other Government stakeholder bodies, such as the Research Councils, can utilise in the formulation of their specific policies.

Within the SIIF, detailed provisions were made for **monitoring and assessment** against a series of deliverables**,** milestones and performance indicators. Thus it was supported by a set of clear targets for achievement on all the dimensions covered, against which the Government and others can track performance over its ten-year implementation period. The Government set out its intentions to publish an annual report on the progress made against various attributes of the science and innovation system (as set out in the framework). A range of indicators was developed against which progress could be judged. In addition, to inform its periodic reviews of public spending (see below), the Government also planned to conduct a detailed assessment of the progress towards the goals for each attribute every two years.

The specific indicators were reviewed by a joint OSI/Treasury steering group in October 2006 and a new reporting framework and set of indicators was developed to assess the health of the science and innovation system, partially replacing and developing upon those used formerly. These fell into a number of categories and 'influence factors':

Categories:

1. Overall economic impacts
2. Innovation outcomes and outputs of firms and governments
3. Knowledge generated by the research base
4. Investment in the research base and innovation

Influence Factors:

A. Framework conditions
B. Knowledge exchange efficiency
C. Demand for innovation

The framework is intended to be used to model the delivery of economic impacts at the aggregate (macro) economy level and it is recognised that "alternative methodologies may be more appropriate (at the micro level) to demonstrate the contribution of key organisations to the delivery of overall economic impact". As an illustration, the category '*Investment in the research base and innovation*', which sets out the landscape of funding for the science and innovation system, will mainly include indicators on:

- Expenditure on R&D, with details of proportions of publicly funded R&D, privately funded R&D, and overseas funded R&D. (Consideration may also be given to the proportion spent on research versus development).
- Other forms of innovation expenditure, as defined by the European Community Innovation Survey.

The process of reporting and review is accompanied by the publication of reports which introduce minor amendments or identify areas for priority action within the 10 Year framework. Thus the strategy process is open to appropriate modifications in the light of evidence-based feedback, informed by a range of stakeholders and reviews. In this way, the impacts of the policy mix can be assessed and adjusted where necessary.

Finally, as noted above, UK government policies are subject to a broader process of review – the Comprehensive Spending Review. The first of these was undertaken in 1997 and a second was launched in 2005, reporting in 2007. The CSR represents a long-term and fundamental examination of Government expenditure, and will shape Departmental allocations for 2008-09, 2009-10 and 2010-11. The CSR programme has involved examination of the key long-term trends and challenges for the next decade (such as demographic and socio-economic change, globalisation, climate and environmental change, global uncertainty and technological change), coupled with a national consultation on how the UK and public services need to respond to these challenges and detailed studies of key areas where the challenges require cross-cutting, innovative policy responses. A set of zero-based reviews of departments' baseline expenditure were also conducted to assess its effectiveness in delivering the Government's long-term objectives. Also within this framework, the Government conducts Spending Reviews which are used to set firm and fixed three-year Departmental Expenditure Limits. These are also subject to PSAs (see above).

1.2 Multiple actors

UK innovation policy formulation involves a number of stakeholders, either directly in the formulation and implementation of policy measures or through a broader process of review and consultation in which many of the actors impacted by policy are involved. The main actors in UK government are HM Treasury, BIS, the Technology Strategy Board (TSB) and the Regional Development Agencies. In this way the views and concerns of the actors responsible for various components of the UK innovation system are able to find a series of common fora and thus the policy mix as a whole can be shaped.

It is also worth noting that a large number of business support schemes are available in the UK. These in turn are operated by a range of organisations at a variety of levels – national, regional and local. This situation is both inefficient and confusing for industry. Consequently, the former Department for Business, Enterprise and Regulatory Reform (DBERR) instigated a major Business Support Simplification Programme (BSSP) to address the problem. Arising from a prior consultation process, the BSSP is intended to provide "a single, cross government oversight of business support, involving a partnership of representatives from all key stakeholders" including the main central government funding departments, Regional Development Agencies and local authorities[142].

1.3 The shift in innovation policy support

Since the late 1970s, the UK has witnessed a clear policy shift away from the direct support of industrial R&D to a broader policy mix of instruments targeting project-based collaborative research between the science base, especially universities, and industry, R&D capacity building in SMEs, the dissemination of innovation expertise, and the development of framework conditions conducive to innovation such as IPR regimes and venture capital and seed financing availability. This has clear consequences in terms of the evaluation of innovation and support and the challenges it faces.

There are two apparent reasons underlying this shift, The first is the realisation that the volume of direct support to private sector R&D from government, in proportional terms, can only be marginal at best and the issue of 'picking winners' – politically an extremely contentious exercise, is also difficult and risky. Moreover, in a conceptual framework influenced by market forces, it is recognised that the rationale for government intervention are strictly bounded by a set of market failure conditions.

The systems view of innovation (as distinct from the linear model viewpoint) also forces attention onto the possibility of system failure rationales for economic underperformance. This further induces government intervention to focus on other (non-private) actors or conditions.

Somewhat paradoxically, however, with the publication of the 1993 Science and Technology White Paper[143] the science base was identified as the major source of innovations that could be readily exploited and developed into new products, processes and services. Since this date and certainly since 1997, the UK Science Base has witnessed a substantial growth in government support for R&D based on what appear to be linear model input-output assumptions. Also of interest is the fact that government policy has effectively undergone a U-turn from regarding R&D tax credits as anathema to the establishment in 2000 of a tax credit scheme for SMEs, with the subsequent extension to large companies two years later. It has to be emphasised that the introduction of the SMEs scheme was the result of an extensive review, by the Treasury, of the use of similar schemes in other countries, whilst an extensive consultation preceded its extension to larger companies. In terms of budget (or, more precisely deferred government income) the tax credit schemes now represent the largest UK policy measure for innovation support, accounting for around 75% of the innovation budget.

---

[142] HM Treasury, Sainsbury Review, 2007

[143] UK Government Realising our Potential: A Strategy for Science, Engineering and Technology. Cm 2250, London, HMSO, 1993.

In parallel, there is increasing policy interest in utilising the, in aggregate terms, enormous amount the Government spends on procurement of products and services to promote private sector innovation. However, progress in this area is largely restricted to debate and the one instrument developed towards this goal, the Small Business Research Initiative, has been somewhat less effective than its US counterpart on which it was modelled, largely due to the need to modify its mode of operation to comply with State Aid restrictions.

2 A strong evaluation culture

The rationale behind the selection of the UK as a focus for this case study is the fact that it is recognised internationally as having a well developed culture of evaluation. The development of this was partially driven by a historical need to apply a greater level of selectivity and prioritisation in the allocation of research funding in the 1970s and 1980s (and to demonstrate efficiency, effectiveness and value-for-money), but it is also coupled to the broader issues of governance such as review and the desire to ensure that policies are appropriate to (and address the issues posed by) the problem for which they have been designed.

In terms of the evaluation of innovation support programmes (including R&D funding programmes) the lead in developing evaluation practice was taken by the DTI, with supporting interest from the Treasury[144] and the National Audit Office, the Government's financial 'watch dog'. DTI also developed its ROAMEF[145] guidance as a tool for programme managers which made the provision of advanced plans for monitoring and evaluation a prerequisite for departmental programme support. Thus evaluation became a strongly entrenched policy tool within DTI and the Research Councils (with a strong academic interest in the topic expressed and supported by the Economic and Social Research Council). Numerous programmes were subject to evaluation, either by dedicated bodies within the funding agencies or by external consultancies.

The specific rationale for selecting programmes for evaluation, according to the DTI's Guidance Plans[146] were:

- To produce information to feed into evaluation against Public Service Agreement, Competitiveness White Paper, Modernising Government, and other key objectives and **targets**;

---

[144] The Treasury published a guide on evaluation to be applied across Government: Her Majesty's Treasury "Green Book" - "Appraisal and Evaluation in Central Government", The Stationary Office, 1997 – see later.

[145] ROAME-F is an acronym for Rational, Objectives, Assessment, Monitoring, Evaluation and Feedback. Each programme, seeking departmental support above a certain financial threshold, was required to have an accompanying ROAME-F statement.

[146] *Guidance on preparing evaluation plans*, ES Central Evaluation Team, DTI, 1999.

- To evaluate programmes, and areas of ‘running cost’ expenditure, with over £10m spend, where there are **significant gaps** in the knowledge of the programmes impact and effectiveness.
- To provide information specifically required for an **identified future need**, for example, programme extension or revision, or possible **replication of a programme in other areas or sectors or on a larger scale**. The specific need should be identified.
- To provide information on the **relative effectiveness** of different programmes, and/or areas of running cost expenditure, in meeting common objectives or on how programmes are complementary.
- To assess the effectiveness of **a new or modified programme or policy.**
- To provide more in-depth assessment of programmes which are **not meeting their objectives and targets.**

In terms of the contribution of evaluations in developing evidence for the formulation of the policy mix, the first, third and fourth bullet points are particularly relevant. Moreover, in defining the rationale for a programme (as part of the ROAME-F approach) the following criteria, as set out in the Guidance, are relevant to the issue of policy mixes (particularly those parts underlined):

- Assess what evidence exists to support the view that there is a market or institutional failure requiring intervention (if applicable);
- Consider the extent to which the reason for intervention is pervasive or limited to certain sectors/firms in the economy;
- Assess the extent that existence of the scheme has impacted on the behaviour of those sectors/firms/market it aims to influence;
- Consider how the relevant market(s) is (are) developing, and possible implications for the continuing validity of the rationale. Account must be taken of the types of activity/projects that the scheme supports. Comparisons must be made with other schemes that support similar activities;
- Identify any links with parallel activities supported by the DTI or by other public and/or private sector bodies, and consider the extent to which the related activities may be complements or substitutes for those being evaluated;
- Consider whether the policy or programme is the right way of tackling any perceived market failure i.e. the rationale may remain valid but there may be more effective mechanisms by which the underlying problems can be tackled.

In addition to the governance processes illustrated above, the results of evaluation can be used in the assessment of which parts (i.e. instruments) of the policy mix are effective, and which are not and how instruments may be modified singly or in combination with others in order to improve the policy mix. Thus, this ‘micro level’ control of the policy mix (in parallel to the macro level control described in the previous section) allows it to develop over time.

3 The historical context shaping evaluation

This short section examines the development of evaluation and appraisal in UK government, with a particular focus on science, technology and innovation policy evaluation.

The UK government’s current policy for evaluation and appraisal can be seen to have experienced a continuous development over nearly a century with the aspects of the current approach going back to at least since the creation in 1914 of the Exchequer and Audit Department (which eventually became what is now known as the National Audit Office or NAO) (Hills and Dale, 1995)[147] but also incorporating important developments brought about within the last few years. The approach of this short review of UK evaluation and appraisal is to focus upon the institutional developments taking place within the UK, of which there have been many, and the corresponding rules, frameworks and approaches to evaluation that such institutional changes have developed and promoted. We explore, briefly, the four major features of this development from the last four decades that have, more than anything else, shaped the current approach to evaluation and appraisal. These are: a) the development of systematic approach to evaluation in the 1970s and 1980s; b) the accumulation of evaluation expertise through limited meta-evaluation that has led to an innovation culture in government which recognises the value of a practical business oriented approach to policy; c) the growing consensus around the neoclassical model of the economy and society; and d) the extension of evaluation activities throughout government as the devolution of policy and programme and project design and their evaluation has been pushed downwards and outwards from Whitehall to the regions.

3.1 Systematic Approaches

In the area of science, technology and innovation policy development and appraisal, a key change to practice came with the introduction of the ROAMEF framework during the 1980s, as government became more aware of the need to make stretched resources effective in achieving impact and therefore of the importance of identifying the basis for government intervention and its accompanying costs and benefits. The ROAMEF framework was mainly the result of Treasury pressure and since its introduction, the majority of major government programmes have been assessed formally against such criteria. ROAMEF evaluation makes major assumptions about rationales, reflecting the influence of economic thinking upon policy evaluation theory and practice, and then works downward from higher to lower level sets of objectives. The ROAMEF framework remains a salient feature of the current approach, but, as we note in the next subsection, important changes have occurred in evaluation as the abstract, programmatic ROAMEF framework has been placed within a new context.

3.2 Learning by Doing – Towards Meta-evaluation

As evaluation took root in the UK civil service, expertise in evaluation in major programme departments such as DTI developed together with an evidence base of a kind that facilitated the comparison of evaluation and appraisal reports and methods. With the rise of critical thinking and reflection about government intervention, its evaluation and findings, particularly within DTI's Central Evaluation Team, there came a realization that the ROAMEF framework was not sufficiently focused on ensuring that interventions were developed with the knowledge of how to make them (and future interventions) effective in practical terms. Accumulated government experience of evaluations and appraisals appears to have gradually convinced those responsible for policy implementation and assessment to develop further criteria against which interventions could be assessed. These criteria addressed, in addition to such matters as rationales, the apparently more prosaic issues of the practicability, relevance, resilience and sensitivity of government interventions.

---

[147] Hills, P.V. and Dale, A.J. (1995) Research and technology evaluation in the United Kingdom, Research Evaluation, Vol. 5, No. 1, pages 35-44.

A number of themes and principles were therefore drawn from the area of business and organisational planning and combined with the existing abstract and high level ROAMEF considerations to provide a new set of criteria for programme evaluation. These are termed the “Business Case for an Evaluation” (see below). Evaluation and appraisal therefore now provide a more extended and complete view of policy, although, as we note elsewhere, while more sophistication in methods may be available, actual resourcing of evaluation in order to obtain data of programme performance is still required to deliver credible evaluations and appraisals.

Meta-evaluation also has other effects upon the innovation process, most notably in terms of greater realism and understanding of what programmes of certain types achieve in terms of their outputs, outcomes and impacts. The large scale review of evaluation studies to better understand additionality of the type carried out by the new Department for Business, Innovation and Skills is a good example of the activities taking place within this new institutional setting for evaluation and appraisal[148].

3.3 Consensus about Economic and Social Models

While Barber[149] wrote sceptically about the role of evaluation in increasing our understanding of the world on which policy acts (Barber 1998; page 9) - “[lacking] a fully specific dynamic long term economic model, a specification of the innovation of economic system whose performance policy makers are trying to improve is impossible” - increasingly during the 1990s and the 2000s, programmes have been targeted and measured (evaluated) against the various and sometimes combined market failures of described by the neoclassical economic model. Such developments could not have taken place without a growing consensus within government and society around this view of the economy. With what amounts to a growing formalization of the policy justification process, a more prescriptive, codified and regimented system of evaluation has emerged. One key aspect of this has been the adoption as a result of work by the OSI and Treasury Steering Group in 2006 of a *categories and influence* factors framework, mainly for understanding policy and its contexts at the macro level.

3.4 Broadening Evaluation and Appraisal in Government

The growth of government programmes, frequent policy change, and above all the decentralization and devolution of policy making to the regional level have significantly increased the number of government and government agency staff involved in the delivery of policy, evaluation and appraisal. For examples of studies of regional evaluation, additionality, programme design and justification for just the London area, see the following: LDA 2006[150], LDA 2009[151], LDA 2005[152], and GLA Economics 2006[153]. To some extent, this has been facilitated by the growth of expertise in central government which has subsequently been available for broader dissemination, but it has also been helped by the use of frameworks in which policy targets and indicators are defined. In relation to additionality, for example, the English Partnerships guide (now in the third edition)[154] and that for Scotland (Scottish Enterprise 2008[155]), represent the key reference documents for policy evaluation and appraisal across the broad range of government activities including science, technology and innovation policy.

---

[148] Department for Business Innovation and Skills, 2009. Research to Improve the Assessment of Additionality. *Rep. No. 1.*

[149] Barber, J. (1998) “Role of Appraisal Monitoring and Evaluation in Policy-Making” in Appropriate Methodological Matrices.

[150] LDA. 2006. The Rationale for Public Sector Intervention in the Economy

[151] LDA. 2009. Business Case Workbook Part 1 Guidance Notes

[152] LDA. 2005. Exploring the case and context for assisting growth businesses in London.

[153] GLA Economics. 2006. The Rationale for Public Sector Intervention in the Economy

Despite the regional character of much recent policy making and evaluation and appraisal activity, a degree of coherence between evaluation approaches across the various levels of government has been maintained through the action of central government bodies such as BIS, the Cabinet Office, and the body set up to promote good programme design and delivery in the regions, the Office of Project and Programme Advice and Training (OffPAT). OffPAT's recent publications promoting good practice in evaluation and policy design include the Office of Project and Programme Advice and Training, 2006[156]) which specifically deals with the use of logic chains in evaluation and appraisal, and its 2008 guide on targets of policy (Office of Project and Programme Advice and Training, 2008[157]).

4 Current practice

4.1 Overview

In 2007, the then Department for Innovation, Universities and Skills (DIUS), the successor to DTI and forerunner of the Department for Business, Innovation and Skills (BIS), assumed general responsibility for all UK innovation activities. In practical terms, this meant that DIUS was required to produce an Annual Innovation Report[158] which detailed government departments' innovation-related activities. Given the broad view of innovation espoused by UK policymakers, it is evident that a wide range of government activities may have an impact on innovation. The Annual Innovation Report was expected to bring together the full set of governmental activities that would contribute to UK innovation overall, both in terms of supporting innovation and developing innovative practices within departments.

However, DIUS (and now BIS) also had oversight of the core range of innovation support policies implemented in England. These might be run in-house or through associated executive agencies such as the Technology Strategy Board, through the Research Councils (particularly in the case of knowledge exchange programmes), or at the regional level through the Regional Development Agencies. Consequently, responsibility for oversight of the evaluation of these innovation support instruments now also resides with BIS.

---

[154] English Partnerships. 2008. Additionality Guide.

[155] Scottish Enterprise. 2008. Additionality & Economic Impact Assessment Guidance Note, A Summary Guide to Assessing the Additional Benefit, or Additionality, of an Economic Development Project or Programme, Appraisal & Evaluation Team, 1st November 2008.

[156] Office of Project and Programme Advice and Training (OffPAT). 2006. Project Advice Note 2/06, A Project Logic Chain (PLC) Approach.

[157] Office of Project and Programme Advice and Training (OffPAT). 2008. Market Failure: Categories and Examples, London

[158] DIUS (2008a): Annual Innovation Report 2008.
http://www.dius.gov.uk/policy/annual_innovation_report.html

4.2 The Green Book

As already noted, general guidance on how UK departments should conduct appraisal and evaluation is provided in the 'Green Book' produced by HM Treasury[159]. This covers a number of policy activities that may be undertaken by government, namely:

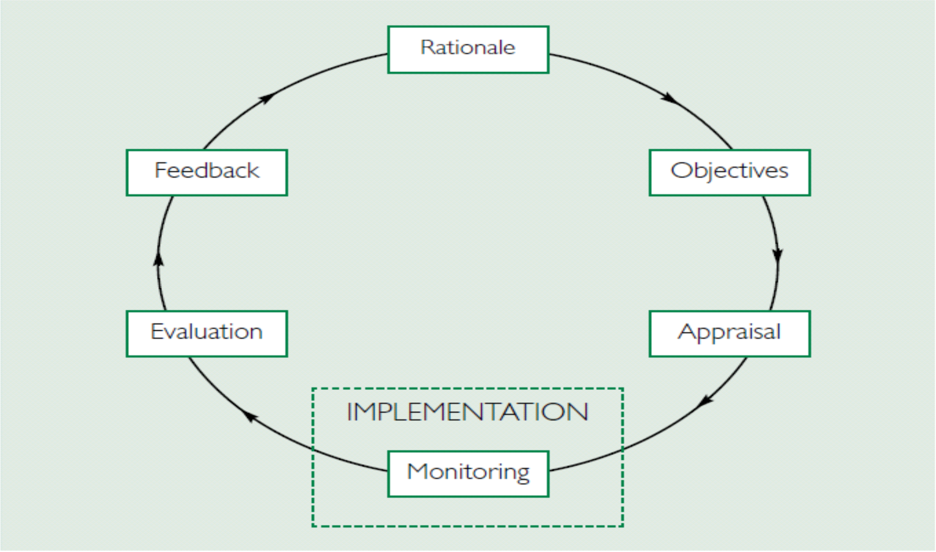
- Policy and programme development
- New or replacement capital projects
- Use of disposal of existing assets
- Specification of regulations
- Major procurement decisions.

Clearly, in the context of this report, the first set of policy activities forms the most relevant.

In essence, the principles underpinned by the Green Book are those which form the primary rationale for the evaluation of UK Government funded projects and programmes, i.e. to ensure that value for money is achieved in promoting the public interest. Its main purpose is "to ensure that no policy, programme or project is adopted without first having the answer to these questions: Are there better ways to achieve this objective? Are there better uses for these resources?". The guidance also emphasises the need to take account of the wider social costs and benefits of proposals, and the need to ensure that public resources are subject to proper use.

While the Green Book offers advice for all central Government departments and executive agencies, it also provides more specific technical advice for specialist analysts and economists in relation to some of the more complex and involved aspects of appraisal and evaluation. Core to the organisation of appraisal and evaluation practice is the aforementioned ROAME-F cycle (see Exhibit 81).

Exhibit 81: The ROAME-F Cycle



(Source: HM Treasury, *The Green Book*)

---

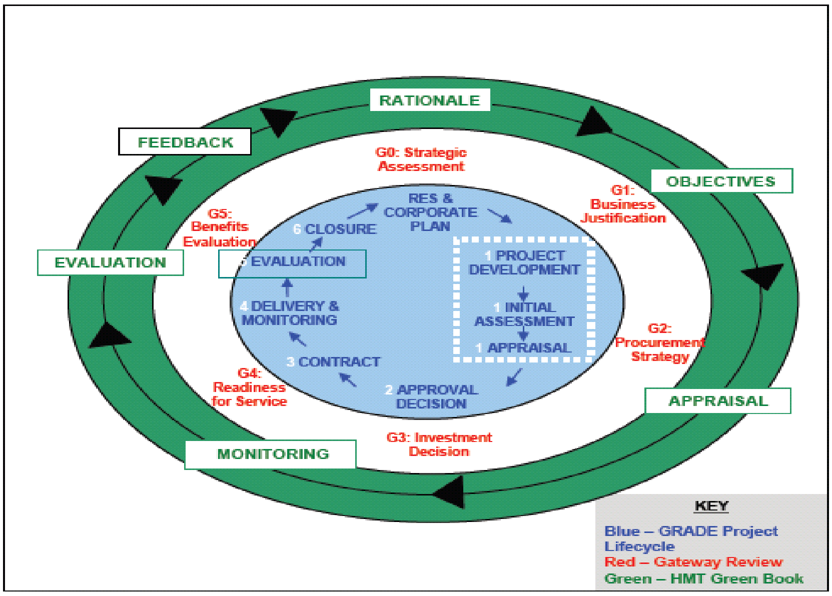
[159] HM Treasury, *The Green Book: Appraisal and Evaluation in Central Government*, Treasury Guidance, London TSO, 2003. http://www.hm-treasury.gov.uk/data_greenbook_index.htm

Against this framework, the Green Book provides guidance on issues such as: the role of appraisal; processes for appraisal and evaluation; setting objectives, outcomes, outputs and targets; and the evaluation process.

The Green Book has also been used to derive more specific guidance for other policy actors in the UK system. For example, in 2008, the then Department for Business Enterprise and Regulatory Reform (BERR)[160], in conjunction with the Regional Development Agencies, English Partnerships and the Office of Project and Programme Advice and Training, produced a *Guidance for RDAs in Appraisal, Delivery and Evaluation.* In part, this was prompted by a policy shift towards a greater level of regional delivery for a number of formerly centrally run innovation support measures. In addition, a number of the innovation support measures that had been transferred to the Technology Strategy Board also took on a greater level of regional engagement. Thus, the regional delivery of innovation support instruments (both those supported by the RDA Single Budgets and by European Funds (e.g. ERDF projects)) and including their appraisal and evaluation, became the responsibility of the Regional Development Agencies in England[161].

The purpose of the *Guidance for RDAs*[162] is to set out the principles and key stages for project development, appraisal, approval, delivery, monitoring and evaluation. It thus responds to the difference in scale between national and regional policy interventions and also links the ROAME-F cycle to both the project life cycle stages and the stages of the so-called Gateway Review[163] of the Office of Government Commerce (see Exhibit 82).

Exhibit 82: Comparison of Project Life Cycle Stages


RATIONALE
FEEDBACK
OBJECTIVES
EVALUATION
APPRAISAL
MONITORING
G0: Strategic Assessment
G1: Business Justification
G2: Procurement Strategy
G3: Investment Decision
G4: Readiness for Service
G5: Benefits Evaluation
RES & CORPORATE PLAN
1 PROJECT DEVELOPMENT
1 INITIAL ASSESSMENT
1 APPRAISAL
2 APPROVAL DECISION
3 CONTRACT
4 DELIVERY & MONITORING
5 EVALUATION
6 CLOSURE
KEY
Blue – GRADE Project Lifecycle
Red – Gateway Review
Green – HMT Green Book

Source: BERR, *Guidance for RDAs,* 2008

---

[160] In 2009, DIUS and BERR were merged to form the Department for Business, Innovation and Skills (BIS)

[161] The Devolved Administrations of Scotland, Wales and Northern Ireland are largely responsible for the delivery of their own innovation support measures and of their own versions of UK-wide support measures.

[162] http://www.berr.gov.uk/files/file45733.pdf

[163] The Gateway Review process is a form of risk assessment procedure applied to a wide range of government acquisition programmes and projects.

It is beyond the scope of this case study to go into a detailed review of the guidance provided by the Green Book and other complementary publications, such as the *Guidance for* RDAs, and interested readers are directed to refer to these directly. In addition, more specific advice and guidance on evaluation for policymakers is provided by the complementary publication – *The Magenta Book* (see below). However, the following extract from the BERR *Guidance* is particularly pertinent, and can also apply at the programme level:

> Evaluation serves two key purposes in the project lifecycle. The principal aim is to assess the extent to which the project has met its original objectives. Part of this process includes testing the assumptions that were made in developing and appraising the project to ensure that lessons are learned for future projects.
>
> All projects should be covered by an evaluation plan as an integral part of project development. This allows mechanisms to be established to record relevant data and ensure learning is captured during the lifetime of the project. The RDA should then ensure that this evaluation plan is implemented, including any necessary amendments to reflect changes in situation. The size, scope and scale of the evaluation(s) planned should be proportionate to the size, complexity, risk and innovation of the project, with relevant resource allocation (time and funding) identified within the evaluation plan.
>
> A final evaluation should be carried out at a period after the project when project impact can be effectively assessed. For larger or more complex projects this is likely to be between 6 months and 5 years after the end of the project. However, it may be important to assess progress towards objectives at an earlier stage e.g. in order to assess effectiveness of the project before allocating continuation funding.
>
> An interim evaluation may then be required within the lifetime of the project. Clear definition of proposed timings and purpose of evaluation should be included in the evaluation plan. These requirements should also be reflected in funding agreements with delivery organisations / funding recipients.
>
> Evaluations should reflect an assessment of the assumptions made in the appraisal process but in the knowledge of what has actually happened in the project. It should focus on how far the critical success factors identified at appraisal were met as well as on resulting outcomes and impact. Lessons learned from evaluation should be fed back into the project process to inform the development, appraisal and delivery of new projects.

It is also worth drawing attention to three specific policy tools that are not referred to directly within the Green Book yet which form a core element in the formulation and implementation of UK government policy support, particularly with regard to appraisal and evaluation: business cases, programme plans and balanced scorecards.

4.3 Business cases, programme plans and balanced scorecards

As is evident from the previous sections, appraisal and evaluation are processes that are firmly embedded in the policy lifecycle of programmes, projects and other policy interventions in the UK.

Whilst it was common practice for programmes of the former DTI to require the completion of a ROAME statement prior to their gaining investment approval, programme managers in DIUS and now BIS, must compile a Business Case.

The Business Case is used “to obtain management commitment and approval for investment in business change including projects and programmes, through rationale for the investment”. In addition, it provides a framework for programme and project planning and management and offers a framework against which the ongoing progress and viability of a project or programme can be monitored. It is prepared by programme management in advance of the design and development of a policy instrument.

The Business Case can be quite an extensive document and addresses a number of headline topics. The initial task is for the compiler to complete a ‘Fitness for purpose checklist’ (e.g. Is the business need clearly stated? Is it clear what will define a successful outcome? Is it clear how the benefits will be realised? etc.)[164]. The Business Case should also contain information covering five key aspects:

- Strategic fit (i.e. description of the business need and its contribution to the organisation's business strategy)
- Objectives (i.e. why it is needed now, the key benefits to be realised and critical success factors and how they will be measured)
- Options appraisal (i.e. high level cost/benefit analysis of (ideally) at least three options for meeting the business need, analysis of 'soft' benefits that cannot be quantified in financial terms, preferred option and any trade-offs)
- Commercial aspects (applicable where there is an external procurement)
- Affordability (statement of available funding and rough estimates of projected whole-life cost of project)
- Achievability (high level plan for achieving the desired outcome, with key milestones and major dependencies, contingency plans, major risks, etc.)
- Source information (procurement documentation, programme/project management plans and documentation, high level requirements, Business Strategy)

Once a policy intervention has been sanctioned, a programme plan is developed. This is used to design the overall programme and then to track and control progress. The programme plan provides a basis for tracking the impact of component projects on the overall goals, benefits, risks and costs of the programme.

The suggested content for the programme plan consists of:

- Project information including the list of projects (the Project Portfolio), their target timescales and costs, and the dependency network showing the dependencies between the projects.
- Summary of risks and assumptions identified against successful completion of the Plan. Detailed assessment of all risks and associated contingency actions is covered in the Risk Register/Log
- Overall programme schedule showing the relative sequencing of the projects, the grouping of projects into tranches, milestone review points.
- Transition Plan showing when the outputs from the projects will be delivered and what transition activities will be required to embed the new capability into business operations.

---

[164] http://www.ogc.gov.uk/documents/BusinessCaseTemplate-MinimalContent.pdf

- Monitoring and control activities, information requirements to support this, performance targets and responsibilities for the reporting, monitoring and control activities[165].

Monitoring of progress and the delivery of the programme against established milestones, targets and deliverables is achieved through the use of a balanced scorecard approach. Periodically, progress is assessed against the balanced scorecard framework and a 'traffic light system' (stop, caution, proceed) is used to indicate the current status and the need for amendment or other interventions to the implementation of the measure.

A full description of this approach is provided in the document *Choosing the Right Fabric: A Framework for Performance Information*[166]. This document, produced by a number of Government bodies concerned with the performance of the public policy system, is intended to offer guidance for public bodies in constructing performance information systems and selecting appropriate performance measures. In its own words, it "should help to spread best practice; establish common principles; and make it easier to integrate national and local performance information systems".

Originally developed in the private sector, the Balanced Scoreboard approach to performance measurement aims to circumvent some of the shortcomings of 'traditional' performance measurement approaches. For instance, by concentrating on sales and profits, performance measurement may appear to indicate success whilst ignoring factors related to the long-term development of a business (such as customer relations or staff training). However, a balanced scorecard approach will measure both the ultimate outcome of the business and aspects of the business that must be maintained in the long term. There is also a balance between financial and non-financial measures and across stakeholders. According to the Treasury document, a balanced scorecard groups performance measures under four headings, namely:

- "The business processes perspective – are the processes within the business working well? Is the organisation producing what it needs?
- The financial perspective – is the organisation operating efficiently and within budget?
- The learning perspective – does the organisation develop its staff, and take on board developments in technology?
- The customer perspective – how do the organisation's customers perceive it? Is the organisation satisfying its main customers?

The groups that are used may vary, according to which factors are important for the success of an organisation. Examples for a government department might include 'meeting the goals of civil service reform' or 'meeting PSA targets'. Of particular relevance to measuring the performance of a programme, the scorecard should be "balanced between the overall objectives of the organisation and the processes and milestones that need to be met in order to achieve these in the short and long term"[167].

---

[165] http://www.ogc.gov.uk/documentation_and_templates_programme_plan.asp

[166] HM Treasury, Cabinet Office, National Audit Office, Audit Commission, Office for National Statistics, *Choosing the Right Fabric: A Framework for Performance Information*, March 2001.

[167] Ibid.

4.4 The Magenta Book

The Magenta Book[168] was produced by the Prime Minister's Strategy Unit in the Cabinet Office in 2003. It has since been updated by the Government Social Research Unit in the Treasury in 2007. It complements other documents such as the Green Book and provides more focused guidance for policy evaluators and analysts, and those who use and commission policy evaluation. In particular, it focuses on policy evaluation in government and it is structured to meet the needs of government analysts and policy makers. It is intended to offer a "user-friendly guide for specialists and generalists alike on the methods used by social researchers when they commission, undertake and manage policy research and evaluation". Moreover, it

> "endeavours to provide guidance on social research methods for policy evaluation in readable and understandable language. Where technical detail is required, or it is necessary to expand on methodological procedures and arguments, these are presented in boxed and shaded areas. It provides examples of evaluations that have used the available methods appropriately and effectively, and it highlights what it is that is good about them. The Magenta Book covers the broad range of methods used in policy evaluation, and the approaches of different academic disciplines (social policy, sociology, economics, statistics, operational research). It is driven by the substantive policy questions being asked of analysts, rather than by methodological disputes between academic disciplines or different schools of thought. It includes guidance on how to use summative and formative, quantitative and qualitative, experimental and experiential methods of policy evaluation appropriately and effectively"[169].

Weighing in at over 200 pages in length, the Magenta Book provides a wealth of information and advice for those interested in evaluation from the full range of policy contexts and is based on extensive government experience in policy evaluation. For example, it addresses questions such as:

- How to refine a policy question to get a useful answer
- The main evaluation methods that are used to answer policy questions
- The strengths and weaknesses of different methods of evaluation
- The difficulties that arise in using different methods of evaluation
- The benefits that are to be gained from using different methods of evaluation, and
- Where to go to find out more detailed information about policy evaluation and analysis

Moreover, it does this at a range of levels from the definition of policy evaluation, through the conduct of meta-evaluations, down to advice on how to word and order questions when designing questionnaires.

Clearly, there is an extensive literature and a range of embedded practices relating to appraisal and evaluation in the UK policy system, all of which reinforces the view that the country possesses a well developed evaluation culture. The scope of the available material precludes any in-depth examination of specific evaluation issues and techniques, thus the following section focuses on a subset of issues that were addressed in the INNO Appraisal and examines the way in which they are approached in the UK.

---

[168] Government Social Research Unit, HM Treasury, *The Magenta Book: Guidance Notes for Policy Evaluation and Analysis*, October 2007.

[169] http://www.gsr.gov.uk/downloads/magenta_book/Intro_Magenta.pdf

5 Specific issues from the survey

The INNO-Appraisal survey examined twelve evaluation reports relating to five UK innovation support measures. This sample size is, thus, too small to enable any statistical analysis to be performed within the subset of UK reports. However, on the basis of qualitative evidence obtained through interviews with UK policymakers, either engaged in evaluation practice or who, as programme managers, commission evaluations, it was possible to examine how some of the specific issues addressed by the survey were dealt with in the UK context.

5.1 Rationale and purpose

As is clear from the above, the primary rationale for the evaluation of innovation support instruments is to ensure that programmes are delivering value for money (i.e. economic benefits), according to HM Treasury principles. However, despite the process of evaluation being standard departmental practice, there is also a very strong element of policy learning involved – has the policy intervention been delivered properly and are there any lessons to be learned from the experience and process. It is worth noting that BIS has a high level evaluation steering group which examines how evaluations are prepared and performed and how evaluation evidence is used in the design and implementation of policy. Related issues typically addressed include the search for both anticipated and unanticipated outcomes, and have there been spill-overs and wider outcomes of relevance to other actors and participants.

5.2 Evaluators

All evaluations have to be contracted out to external evaluators. BIS has no in-house competence or expertise to conduct evaluations, although staff are fully aware of the issues and possible approaches to evaluation (not least because of the range of supporting documentation and training available within government). In addition, there is a requirement that programmes should be evaluated by those independent of the programme – thus in the absence of in-house resources for a full evaluation, external consultants tend to form the favoured option. Generally, an open tender procedure is used in order to invite evaluation proposals. One of the key issues in the selection process is that evaluators have to demonstrate a strong understanding of the programme, what it is trying to deliver and what lessons and feedback the programme management are asking the evaluation to deliver. Obviously, a demonstrable track record and evidence of competence and relevant expertise are also crucial. See the OGC website for further information[170].

5.3 Terms of Reference and the use of innovative approaches

The Terms of Reference for an evaluation are set in house, based on internal advice. Generally, Terms of Reference are set according to the Green Book principles and Business Plan templates. Terms of Reference are quite prescriptive in order to guarantee that the purposes of the evaluation are met; nevertheless, programme managers would be open to the use of new, exploratory or innovative evaluation approaches in addition to those accepted as more ‘traditional’ approaches. Since time pressures are highly relevant for the delivery of an evaluation, there tends to be little opportunity for evaluations to involve more experimental approaches unless they are able to guarantee the delivery their main intended outcome.

---

[170] http://www.ogc.gov.uk/documentation_and_templates_tendering_process_for_consultancy_support.asp

5.4 Timing

As already noted above, there is extensive use made of monitoring approaches in the implementation of programmes. Thus, issues concerning programme management, the level of uptake, problems arising with the delivery of the programme and similar issues can be detected during the programme life cycle. As most UK innovation policy instruments tend to be of a rolling rather than finite nature, then evaluations tend to be conducted at periodic intervals and are rarely *ex post*. Obviously, where programmes have been introduced to deal with a short term problem or where they have been subsumed into complementary or successor programmes, some form of *ex post* evaluation will be conducted. Thus, timing of evaluations is essentially dependent on the life cycle of the programme in question.

5.5 Conditionality

In a sense, all evaluations of innovation support measures are implicitly conditional according to HM Treasury requirements. As HM Treasury is the ultimate sponsor for policy interventions, then this could be perceived as a sponsor-imposed condition. However, as has been demonstrated, evaluation culture is developed to such an extent in the UK that the notion of conditionality, as interpreted in the INNO Appraisal survey, is not applicable. Similarly, whilst the evaluation of ERDF funding, for example, is a pre-condition, it does not in itself form the sole demand for evaluation of supported programmes.

5.6 Dedicated budget for evaluation

As evaluations are foreseen and planned under the ROAME-F and Green Book frameworks, a budget for evaluation is foreseen during the planning stage. As a broad rule-of-thumb, the budget will be around 0.5% of the programme budget. However, this greatly depends on the scope or scale of the programme/instrument (a small programme may require proportionately more budgetary resources to enable a meaningful study to be conducted, whilst a very large programme might be serviced by a comparatively modest evaluation). It would also depend whether the programme is a new one or if there is already accumulated evidence and existing evaluations which can contribute to the foreseen evaluation.

5.7 Is evaluation foreseen and planned?

As noted under the previous point, in the UK all innovation programmes are developed and formulated in such a way that appraisal, monitoring and evaluation are all anticipated elements within the programme life cycle.

5.8 Topics, data collection methods, and data analysis methods

Each of these aspects is dependent on the specific nature of the innovation support instrument being evaluated and on the objectives of the evaluation. Clearly, the topics selected for coverage by the evaluation will, to a large extent, dictate the precise data collection methodologies to be employed, which in turn will determine the data analysis methods. The Magenta Book[171] provides a thorough treatment of a wide range of evaluation approaches and methodologies, offering rationales for their use and advice on their application, advantages and disadvantages, and gives a picture of those that are typically available to UK evaluators and policy analysts.

---

[171] http://www.gsr.gov.uk/downloads/magenta_book/Intro_Magenta.pdf

5.9 Programme impacts

Again the types of impact, either anticipated or unanticipated, will depend upon the nature of the innovation support intervention. Evidence from BIS policy makers indicates that the objective of programme evaluations is to seek for both forms of impact. Clearly, the need to demonstrate that a programme has achieved value for money implies that the determination of economic impact will often be an evaluation goal, although this may not always be the explicit objective of the programme (for example, raising firms' awareness of innovation potentials, increasing science-industry knowledge exchange). The Magenta Book provides extensive discussion on the nature of programme impact and ways in which it may be measured[172].

5.10 Sponsors, audiences and availability of results

As already noted (Section 5.5), the ultimate sponsor of UK innovation policy support tends to be HM Treasury, although European funding is also received, e.g. via the ERDF. However, the audience for evaluation reports is frequently wider. The Treasury does have an interest and oversight for the outcomes of programme evaluation, but, as also mentioned, evaluation implementation and practice is also scrutinised by a high-level steering group in BIS. Other potentially interested parties for evaluation outcomes could include the National Audit Office, and Parliamentary committees, such as the Public Accounts Committee and advisory committees in the House of Commons and the House of Lords. In addition, policy analysts (from both the private and public sectors) form a further, albeit rather restricted, potential audience.

In general, all BIS evaluation reports are made publicly available, the more recent ones being published on the web[173]. Similarly, the reports of evaluations conducted by government agencies such as the Technology Strategy Board and the Regional Development Agencies may be found on the relevant websites[174]. If a report runs the risk of breaking confidentiality, i.e. by identifying specific participants or discloses sensitive company information, then it may not be published or only published in part.

5.11 Recommendations

In general recommendations are acted upon, provided they meet the conditions of being realistic and economically feasible. It is not possible to provide a more specific indication of the extent to which recommendations are followed up as this depends on the nature of individual programmes and of the recommendations themselves. Generally, if a report contains recommendations, these will also be published unless (as in the case of publication of evaluation reports above) there is a risk that doing so will infringe business confidentiality (i.e. it is possible to identify specific participants, such as large firms) or they include sensitive disclosures.

5.12 Quality and utility

The quality, of an evaluation, is defined by BIS policymakers as being fit for purpose, i.e. the evaluation meets the Terms of Reference within a reasonable budget and also delivers recommendations that are feasible and realistic. It was agreed that quality is an asymptotic function: that is there is a minimum level of quality that must be achieved for the delivery of the evaluation's objectives – i.e. the programme manager who commissioned the evaluation must have confidence in the validity of the results and the recommendations. Any increase in the level of quality (i.e. through more complex data collection techniques, elaborate forms of analysis, etc,) incurs a law of diminishing returns (in terms of the usefulness of the evaluation).

---

[172] Ibid.

[173] http://www.berr.gov.uk/publications/economicsstatistics/economics-directorate/page21981.html

[174] For example, see: http://www.nwda.co.uk/search-results.aspx?terms=evaluation%20&btnSubmit=Go&

Similarly, an evaluation is deemed to be useful if the evaluation delivers the Terms of Reference in a consistent manner and if it provides actionable recommendations and delivers value for money. Usefulness can be defined as the degree to which there is feedback on policy and if something was learned from the process of the evaluation. However, the timing of an evaluation could have an impact on its usefulness – too early in the programme life cycle and there would be little to be learned, too late and it would not be possible to put the policy lessons into effect.

6 Conclusion

It is difficult to provide a set of conclusions from this particular case study as its underlying rationale was to provide evidence on the existence of a culture of evaluation within the UK.

It is clear that the identification of the UK as a leading exponent of innovation policy evaluation has been substantiated. There is an extensive range of policy literature which offers guidance on the practice of evaluation and appraisal within programme management, which has been developed through years of experience in the field of innovation policy evaluation. Tailored support for evaluation is offered at a range of government levels and there is a strong government-wide imperative for its conduct.

However, the availability of supporting literature alone cannot be a substitute for experience in the use and practice of evaluation – it can only serve to complement and introduce the necessary concepts to a new audience. Thus, in addition, this guidance is not merely prescriptive but is supported with more detailed advice which seeks to engage policy makers in a wider and more considered appreciation of the approaches and uses of evaluation (as part of a broader system of performance measurement) and to develop better understanding of the principles and methods involved.

To develop a fully evaluation culture requires not only the availability of practical and appropriate methodologies and approaches but also an understanding of the relevance and policy benefits that can accrue from their application. It is clear that the use of evaluation as a learning tool for innovation policy support is thus well-established and understood as such in the UK.

The development of evaluation and appraisal in government in the UK is thus a story of increasing use of the practice of evaluation, to justify the allocation of resources, to quantify outputs, outcomes and impacts in relation to a defined set of policy targets.

PRO INNO EUROPE
INNO
APPRAISAL

Part III Chapter 11

The case of the Mediterranean Countries (Cyprus, Greece, Italy, Malta, Portugal and Spain)

The focus of this study is to examine the current situation in the six Mediterranean countries (Cyprus, Greece, Italy, Malta, Portugal and Spain) with regards to the ways evaluations are carried out. It is based on the results of the specific questionnaire survey carried out under the INNO-Appraisal study and more specifically, the evaluation topics covered, the identified data analysis and collection methods, as well as the quality and usefulness of the evaluations. These results are then compared to the overall results of the INNO-Appraisal study so as to identify any possible inconsistencies and differences.

Given that the evaluations in the countries under focus are governed mainly by Structural Funds (SF) requirements, the results are similar to those of the Structural Funds type evaluations examined in chapter 7. Nonetheless, this study is drawn on an entirely different evidence base (all SF type evaluations in chapter 7 vs. the six countries in this section). Yet, the results are supporting evidence to the conclusions of chapter 7, i.e. SF regulations do make a difference in terms of how the evaluation types are conducted, but do not seem to indicate anything different from what is usually dictated by international practice, which is reflected in the overall results as well. Another interesting conclusion is that the quality of evaluations carried out in the **six** countries, although lower in comparison with the overall results, is between 3 and 4 on an 1-5 point scale. This can be considered as a positive impact of SF regulations especially considering the lack of evaluation traditions in these countries. However, despite the relatively good quality of these evaluations, their results are rarely discussed with government cycles or relevant stakeholders, which is another striking difference with the overall results.

Effie Amanatidou and Ioanna Garefi[175]

---

[175] The lead authors for this section of the report are as given above, with contributions from all other team members at different parts of its production, conclusions and interpretations have been approved and are shared by the group.

Table of Contents

Table of Contents 308

List of Tables 309

Table of Exhibits 309

Executive Summary 310

1 Introduction 311

2 Major results of the data collected 311

  2.1 Evaluation Type and Evaluation Topics covered 313

  2.2 Data Analysis Methods covered 316

  2.3 Data Collection Methods covered 318

  2.4 Impacts covered and Main audiences 320

  2.5 Quality of evaluations in the Mediterranean countries 321

    2.5.1 Distributions across quality categories 322

List of Tables

Table 106: Evaluation Topics covered....314
Table 107: Evaluation Topics covered across the different evaluation types 316
Table 108: Data analysis methods covered across the two samples....317
Table 109: Data analysis methods covered across the different evaluation types 318
Table 110: Data collection methods covered across the two samples....318
Table 111: Data collection methods covered across the different evaluation types 319
Table 112: Intended Audiences covered 321

Table of Exhibits

Exhibit 83: Evaluation types across the two groups 313
Exhibit 84: Impacts covered in Mediterranean countries 320
Exhibit 85: Quality of evaluations across evaluation types in Mediterranean group....322
Exhibit 86: Distribution across quality characteristics with the Mediterranean group....323
Exhibit 87: Distribution across usefulness categories within the Mediterranean group 324
Exhibit 88: Dissemination of evaluation results....325
Exhibit 89: Consequences of evaluations in the Mediterranean group 325

Executive Summary

The aim of this case study is to examine the present situation in the six Mediterranean countries (Cyprus, Greece, Italy, Malta, Portugal and Spain) with regards to the ways evaluations are carried out. It is mainly based on the results of the specific questionnaire survey carried out under the INNO-Appraisal study and more specifically focuses on the evaluation topics covered, the identified data analysis and collection methods, as well as the level of quality and usefulness of the evaluations. These findings are then compared to the overall results of the INNO-Appraisal study in order to examine possible identified inconsistencies and differences.

Given that the evaluations in the countries under the focus of this study are mainly carried out according to Structural Funds requirements, the results are similar to those of the Structural Funds type evaluations examined in chapter 7. However, the evidence base is different; all SF type evaluations in chapter 7 compared with non SF type evaluations vs. the six countries' results compared with the total results of the INNO-Appraisal survey.

The initial research hypotheses were as follows:

- Specific evaluation topics are covered in the countries examined vs. the overall results;
- Specific data analysis and collection methods are followed in these countries;
- Specific audiences are addressed;
- Specific quality characteristics are covered;
- Specific issues of usefulness, dissemination and consequences are addressed.

The specific case study mainly draws upon the results of the specific questionnaire template survey carried out under the INNO-Appraisal in comparison with the overall results of the project in order to discover differences and draw substantial conclusions, as well as test whether the aforementioned hypotheses made are indeed the case in the Mediterranean group of countries.

The survey has indicated a small number of differences, but mainly across the different evaluation types, rather than across the Mediterranean countries and the overall population. This suggests that what really makes the difference is SF regulations in terms of how the evaluation types are conducted, but do not seem to suggest anything different from what is usually dictated by international practice, something which is also reflected in the overall results. In terms of quality characteristics, all of them are less satisfied in the case of the Mediterranean countries in comparison with the overall results (as presented in section 3). Yet, when examining the results in the six countries in isolation, it is interesting to note that almost all quality characteristics score between 3 and 4 on a 1-5 point scale in terms of satisfaction. This fact can be considered a relatively positive impact of SF regulations given the lack of evaluation tradition in these countries . However, despite the relatively good quality of these evaluations, their results are rarely discussed with government cycles or relevant stakeholders, which is another striking difference with the overall results.

1 Introduction

The Mediterranean countries under examination in the present study share a common generic characteristic. Albeit to differing degrees, their national and regional research and innovation policies are strongly dependent on Structural Funds. Monitoring an evaluation carried out in accordance with Structural Funds regulations and design specifications, has had a positive impact in some cases. As the Greek Country Report (INNO-Policy TrendChart, 2007) indicates, involvement in Structural Funds accelerated the introduction of evaluation in all levels of programme funding. The Portuguese Country Report (INNO-Policy TrendChart, 2007) shows that the preparation of the National Strategic Reference Framework and its accompanying Operational Programmes, had a greater level of influence on enhancing policy-making coordination compared to the Lisbon National Reform Programme.

However, an ‘evaluation tradition’ especially for policy evaluation is still far from being established, especially in newer Member States, but the situation seems to be changing. The role of external, independent evaluations is increasingly gaining recognition and new mechanisms and agencies are being set up. For example, a system for monitoring and evaluation (SESI) was established under the Ministry of Industry, Tourism and Commerce in Spain, as well as the National Agency for Evaluation of Public Policies & Quality of Services (AEVAL). A new agency (ANVUR) for the evaluation of research results by universities and research organisations was also set up in Italy, in addition to the National Innovation Agency directed towards the evaluation of innovation projects.[176]

2 Major results of the data collected

Preliminary findings are based on 46 completed (out of which, 24 validated) evaluation templates. The majority (26, out of which 19 validate) regards templates for Greek relevant measures, and thus a certain level of bias has to be kept in mind.[177]

A very important trait shared among these countries, mainly owing to their dependence on Structural Funds procedures, is that most evaluations are ‘portfolio’ type. Interim and ex-post evaluations are usually carried out at the level of the Operational Programme’s ‘measure’, which includes several innovation actions or programmes. To illustrate this with an example, the 26 templates that were filled in for evaluations of relevant Greek measures correspond to only 9 individual evaluation documents, two of which already cover 17 out of the 26 measures (14 and 3 respectively).

The findings presented in this chapter are mainly based on the following cases, showing the number of evaluation templates filled in as well as the countries in which the responsible project manager has validated the respective results:

- **Sample A (INNO-Appraisal survey questions C) - Mediterranean countries:** 22 cases (results for Cyprus, Spain, Greece, Malta, Portugal and Italy).

---

[176] INNO-Policy TrendChart — Policy Trends and Appraisal Report, ITALY, 2008

[177] This is mainly because the templates corresponding to the Greek evaluations include both templates for evaluation of the overall programme (portfolio evaluations 4) as well as of certain individual measures under the ‘umbrella’ programmes. This is not the case for the other countries where the corresponding templates mainly refer to evaluations of the overall programmes (portfolio evaluations).

- **Sample B (INNO-Appraisal survey questions D-F) - Mediterranean countries:** 24 cases (results for Cyprus, Spain, Greece, Malta and Italy since no validation template has been received from the Portuguese project managers).

A first indication prior to the conduction of this survey and its validation was mainly based on the fact that most of the evaluations recorded are interim and ex-ante. This was considered normal, given the period covered by INNO-Appraisal and Structural Funds procedures. The study covers any evaluation conducted since the year 2000. This means that the available ones are in fact the ex-ante and interim evaluations for the previous programming period (i.e. around 2000-2006)[178], as well as the ex-ante evaluations for the next programming period (i.e. around 2007-2013). Ex-post evaluations drawn in accordance with Structural Funds procedures and specifications were not available yet as they have to be carried out after the end of the programming period, which will be concluded by the end of 2008 with the application of 'n+2' rule.

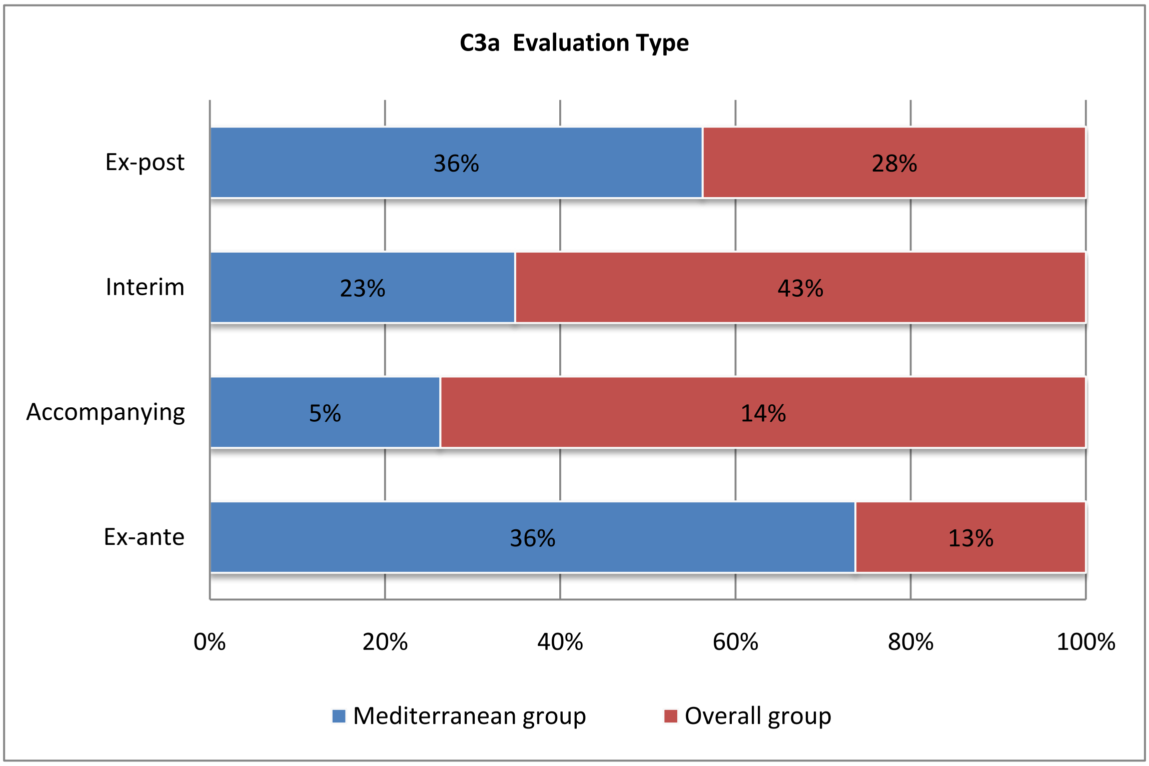
However, the results of the survey have revealed a level of difference in this respect. As seen in Exhibit 83, in their majority, evaluations conducted in the Mediterranean countries are of ex-ante and ex-post nature (36% respectively), while interim evaluations occupy a share of 23%. This greater number of ex-post evaluations compared to interim evaluations is mainly owing to the fact that recorded evaluations in Spain and Italy are of ex-post nature (2 and 5 respectively out of 8 in total). The majority of the evaluations recorded were contracted to external bodies by means of open tendering procedures. A very important issue that needs to be recorded is the fact that Mediterranean countries do not present significant shares in conducting 'Accompanying evaluations', when compared to the overall results of the INNO-Appraisal survey.

This is something normal though, given the nature of Mediterranean evaluations which are drawn in accordance with SF regulations. SF regulations prescribe the conduct of ex-ante, ex-post and interim or mid-term evaluations. The latter type of evaluations is usually carried out at one point during the life time of the programme, rather than multiple points. Thus, they do not qualify as accompanying evaluations.

---

[178] Slight variations in programming periods exist across countries.

Exhibit 83: Evaluation types across the two groups



2.1 Evaluation Type and Evaluation Topics covered

There appear to be less striking differences than one would expect between the two groups, as far as the evaluation topics covered. In fact the issues of ‘internal / external consistency’, ‘goal attainment/effectiveness’, ‘outputs, outcomes and impacts’, ‘programme implementation efficiency’ and ‘policy / strategy development’ are covered by significant shares in both cohorts.

‘Behavioural additionality’ is slightly more addressed in the overall results of the study, whilst gender and minority issues are analysed in more detail in the results of the Mediterranean countries’ study, given also the fact that they form explicit sections in the relevant documents under SF regulations which must be filled in and submitted.

Table 106: Evaluation Topics covered

	Mediterranean Countries		Overall Sample	
Evaluation Topics	Count	%	Count	%
C.7.a External Consistency	18	100,00%	128	78,50%
C.7.b. Internal Consistency	17	100,00%	132	80,50%
C.7.c. Coherence/Complementarity	17	94,40%	96	62,30%
C.7.d. Goal Attainment/Effectiveness	10	71,40%	141	89,20%
C.7.e. Outputs, Outcomes and Impacts	17	85,00%	151	91,50%
C.7.f. Quality of Outputs	7	53,80%	86	57,30%
C.7.g. Value for Money/RoI/Cost-Benefit	4	33,30%	39	27,30%
C.7.h. Programme Implem. Efficiency	12	85,70%	118	76,10%
C.7.i. Project Implementation Efficiency	8	61,50%	71	47,00%
C.7.j. Input Additionality	9	56,30%	74	50,00%
C.7.k. Output Additionality	9	56,30%	75	50,00%
C.7.l. Behavioural Additionality	8	44,40%	74	50,30%*
C.7.m. Policy/Strategy Development	11	64,70%	125	76,20%
C.7.n. Gender issues	10	58,80%	36	23,80%
C.7.o. Minority issues	7	41,20%*	10	6,90%

It might be expected that there are more distinct differences in the evaluation topics covered when looking at the different evaluation types (ex-ante, interim and ex-post) , given the fact that they follow certain instructions on how to conduct these evaluations under the SF framework. Indeed, some minor but clearer differences do appear when looking at the different evaluation topics examined in both cohorts across the different evaluation types as highlighted in Table 107.

Ex-ante evaluations are 'light' and general types of evaluations referring to the whole Operational Programme with no specific references to any measure or programme. They are usually concerned with the **external / internal consistency** and **coherence** of the programme but also with **policy development** and **gender / minority issues**. An important result worth referencing is that apart from all those topics examined and according to the overall results of the INNO-Appraisal survey, outputs, outcomes and impacts are also being addressed presenting a high share among the different evaluation topics examined.

Interim evaluations are more monitoring reports mainly measuring progress against the indicators set, and absorption of funds, but also making suggestions for restructuring of certain measures. They too consider issues of consistency, coherence and complementarity, but also goal attainment, outputs and results as well as programme implementation efficiency. On the other hand, as presented in the overall results, interim evaluations mainly refer to the same issues with coherence and complementarity issues gathering lower shares of preference than those recorded in the Mediterranean countries results. Additionally, policy and strategy development issues are covered in the overall results sample by significant larger shares than the Mediterranean population.

The ex-post evaluations recorded in the Mediterranean countries sample mainly consider issues of effectiveness, outputs and results, quality of outputs, efficiency as well as they present a significant share of input and behavioural additionality. The overall results mainly refer to the same issues with 'external consistency' presenting a vey high share together with 'programme implementation efficiency'.

A concluding remark could be attributed to the fact that for each evaluation type across the two samples (e.g. ex-ante Med countries vs. ex-ante overall results) the picture is rather uniform and there are a lot of commonalities presented with only some differences not that striking in the issues covered. Thus, evaluation topics covered in the Mediterranean countries are not that different from those covered by the total group of respondents for each evaluation type.

This result is reinforcing the result of the case of SF vs. non SF type evaluation results as presented in section 4.3. The differences in the evaluation topics covered by each evaluation type within the Mediterranean group can be attributed to the SF regulations regarding the focus that each evaluation type should have. The fact that the results are very similar in comparison with the overall results reinforces the conclusion of the SF vs. non SF case, i.e that SF regulations suggest what is usually done in international practice.

Table 107: Evaluation Topics covered across the different evaluation types

	Mediterranean countries				Overall results			
Evaluation topics	Ex ante	Acc	Interim	Ex post	Ex ante	Acc	Interim	Ex post
C.7.a External Consistency	44,44%	5,56%	27,78%	22,22%	85,71%	83,33%	76,71%	64,58%
C.7.b. Internal Consistency	41,18%	5,88%	29,41%	23,53%	95,24%	87,50%	80,82%	58,33%
C.7.c. Coherence/Complementarity	44,44%	5,56%	27,78%	16,67%	80,95%	58,33%	53,42%	50,00%
C.7.d. Goal Attainment/Effectiveness	0,00%	7,14%	35,71%	28,57%	38,10%	95,83%	91,78%	81,25%
C.7.e. Outputs, Outcomes and Impacts	25,00%	5,00%	25,00%	30,00%	71,43%	100,00%	90,41%	87,50%
C.7.f. Quality of Outputs	7,69%	7,69%	7,69%	30,77%	23,81%	70,83%	41,10%	64,58%
C.7.g. Value for Money/RoI/Cost-Benefit	16,67%	-	0,00%	16,67%	9,52%	20,83%	19,18%	35,42%
C.7.h. Programme Implem. Efficiency	28,57%	-	35,71%	21,43%	57,14%	70,83%	72,60%	66,67%
C.7.i. Project Implementation Efficiency	7,69%	-	23,08%	30,77%	14,29%	45,83%	41,10%	50,00%
C.7.j. Input Additionality	6,25%	-	18,75%	31,25%	28,57%	45,83%	35,62%	60,42%
C.7.k. Output Additionality	6,25%	6,25%	18,75%	25,00%	28,57%	58,33%	34,25%	58,33%
C.7.l. Behavioural Additionality	0,00%	5,56%	5,56%	33,33%	19,05%	54,17%	39,73%	56,25%
C.7.m. Policy/Strategy Development	29,41%	-	17,65%	17,65%	85,71%	83,33%	76,71%	58,33%
C.7.n. Gender issues	41,18%	-	11,76%	5,88%	57,14%	37,50%	16,44%	2,08%
C.7.o. Minority issues	35,29%	-	0,00%	5,88%	33,33%	8,33%	0,00%	2,08%

2.2 Data Analysis Methods covered

Both samples do not present any striking differences, in terms of the data analysis methods used. When the findings are examined across the two samples we have ‘descriptive statistics’ and ‘context analysis’ standing out in both cases. ‘Document analysis’ comes next in either case, while in the overall results the use of ‘case study analysis’ features a significant share compared to the Mediterranean population.

Table 108: Data analysis methods covered across the two samples

	Mediterranean Countries		Overall Sample	
Data Analysis Methods	**Count**	**%**	**Count**	**%**
C.12.a. Case Study Analysis	3	17,60%	68	41,50%
C.12.b. Network Analysis	1	9,10%	26	17,30%
C.12.c. Econometric Analysis	8	42,10%	36	22,80%
C.12.d. Descriptive Statistics	18	85,70%	121	75,60%
C.12.e. Input/Output Analysis	8	47,10%	41	25,90%
C.12.f. Document Analysis	15	83,30%	82	51,60%
C.12.g. Context Analysis	18	100,00%	106	67,10%
C.12.h. Before/After Group Compar.	6	40,00%	16	10,30%
C.12.i. Control Group Approach	6	40,00%	30	20,00%
C.12.j. Counter-Factual Appr.	3	23,10%	32	21,90%
C.12.k. Cost/Benefit Approach	8	50,00%	36	22,90%

When looking at data analysis methods which are used in the different types of evaluations (ex ante, accompanying, interim and ex post) there are some minor but not distinct differences in both cohorts.

Interim evaluations depend on context analysis in a review mode, descriptive statistics, and document analysis. **Ex-ante** evaluations seem to depend on data analysis methods such as ‘context analysis’, ‘document analysis’, ‘cost/ benefit approach’ and ‘descriptive statistics’ coming next. Within the Mediterranean sample, it is ‘document analysis’ and ‘context analysis’, which are dominant in 2 out of 4 evaluation types **(ex ante and interim)** whereas in the overall results the same picture emerges, even in the ex-post type and accompanying evaluations.

Ex-post evaluations use mainly ‘descriptive statistics’ in both cases but, the Mediterranean group prefers ’Econometric analysis’ or ‘Control group approach’ over ‘document analysis’, which is mainly featured in the overall results group together with ‘context’ analysis. Moreover, ‘case study’ analysis presents a very high share, as expected, within the overall results sample across interim and ex-post type evaluations, whereas in the Mediterranean results this topic presents very low and inexistent shares in the different evaluation types.

As far as the **accompanying** evaluations are concerned there are no striking results within the Mediterranean countries group, since it is not that preferable, whereas in the overall results group the most dominant data analysis methods identified are ‘descriptive statistics’ coming first, followed by ‘context analysis’.

Table 109: Data analysis methods covered across the different evaluation types

	Mediterranean countries				Overall results			
Data Analysis Methods	Ex ante	Acc	Interim	Ex post	Ex ante	Acc	Interim	Ex post
C.12.a. Case Study Analysis	0,00%	-	11,76%	5,88%	4,55%	50,00%	49,32%	37,50%
C.12.b. Network Analysis	0,00%	-	0,00%	9,09%	0,00%	20,83%	10,96%	25,00%
C.12.c. Econometric Analysis	0,00%	5,26%	10,53%	26,32%	9,09%	33,33%	16,44%	27,08%
C.12.d. Descriptive Statistics	23,81%	4,76%	23,81%	33,33%	36,36%	83,33%	80,82%	64,58%
C.12.e. Input/Output Analysis	5,88%	-	23,53%	17,65%	18,18%	41,67%	17,81%	27,08%
C.12.f. Document Analysis	44,44%	5,56%	27,78%	5,56%	72,73%	37,50%	47,95%	41,67%
C.12.g. Context Analysis	44,44%	5,56%	27,78%	22,22%	77,27%	79,17%	52,05%	60,42%
C.12.h. Before/After Group Compar.	13,33%	6,67%	6,67%	13,33%	18,18%	25,00%	1,37%	8,33%
C.12.i. Control Group Approach	13,33%	-	0,00%	26,67%	9,09%	8,33%	17,81%	27,08%
C.12.j. Counter-Factual Appr.	0,00%	-	0,00%	23,08%	0,00%	12,50%	15,07%	37,50%
C.12.k. Cost/Benefit Approach	31,25%	-	0,00%	18,75%	31,82%	29,17%	10,96%	27,08%

2.3 Data Collection Methods covered

Data collection methods as presented in Table 111 show commonalities in the results between the two groups. ‘Existing surveys/ databases’, ‘participant surveys’ as well as ‘document search’ and ‘monitoring data’ are mainly covered in both cohorts with ‘Interviews’ showing a very large share in the overall results sample.

Table 110: Data collection methods covered across the two samples

	Mediterranean Countries		Overall Sample	
Data Collection Methods	**Count**	**%**	**Count**	**%**
C.13.a. Existing Surveys/Databases	21	95,50%	115	69,70%
C.13.b. Participant Surveys	14	70,00%	105	64,80%
C.13.c. Non-participant Surveys	5	33,30%	36	24,80%
C.13.d. Interviews	10	52,60%	128	76,20%
C.13.e. Focus Groups/Workshops/Meetings	7	50,00%	79	49,70%
C.13.f. Peer Reviews	1	7,70%	27	18,60%
C.13.g. Technometrics / Bibliometrics	1	8,30%	3	2,00%
C.13.h. Document Search	17	94,40%	102	64,20%
C.13.i. Monitoring Data	13	81,30%	119	79,30%

When looking at the different evaluation types (ex-ante, interim and ex-post) in terms of **data collection** methods, the situation is similar and even more uniform.

In the Mediterranean group, ex-ante evaluations seem to depend on data collection methods like document search, monitoring data, existing surveys, and workshops. In ex-post type evaluations 'participant surveys' and 'existing surveys/ databases' are the most dominant. 'Monitoring data' and 'document search' feature in 2 out of 4 evaluations types (ex ante and interim). Differences refer mainly to the increased use of 'Existing surveys/databases' and 'focus groups and meetings' in the ex-ante evaluations whereas 'Interviews' indicate very low shares of significance in all four types.

The other point of difference refers to the increased use of 'participant surveys' in interim evaluations which is again expected as this is the usual method applied in evaluating the progress, as well as the outcomes and impacts of interventions. Other than that it is the 'existing surveys', 'document search' and 'monitoring data' following that are mostly used in all evaluation types in the Mediterranean group.

The situation is not at all different in the overall results group where 'monitoring data', 'existing surveys' and also 'interviews' feature as highly significant across all types of evaluations, while 'participants surveys' in 3 out of 4 evaluation types. Examining the findings under each evaluation type across the two groups, the main difference is the additional preference for 'participant surveys', 'interviews' and 'document search' for the ex-post type evaluations of the overall results group.

These results are similar to the SF vs. non SF type evaluations as presented in section 4.3. Differences in the data analysis and collection methods are more marked across different evaluation types rather than across the Mediterranean countries and the overall population. This suggests that SF regulations do make a slight difference but do not seem to suggest anything different than what is usually dictated by international practice for the different evaluation types and thus appearing in the overall results as well.

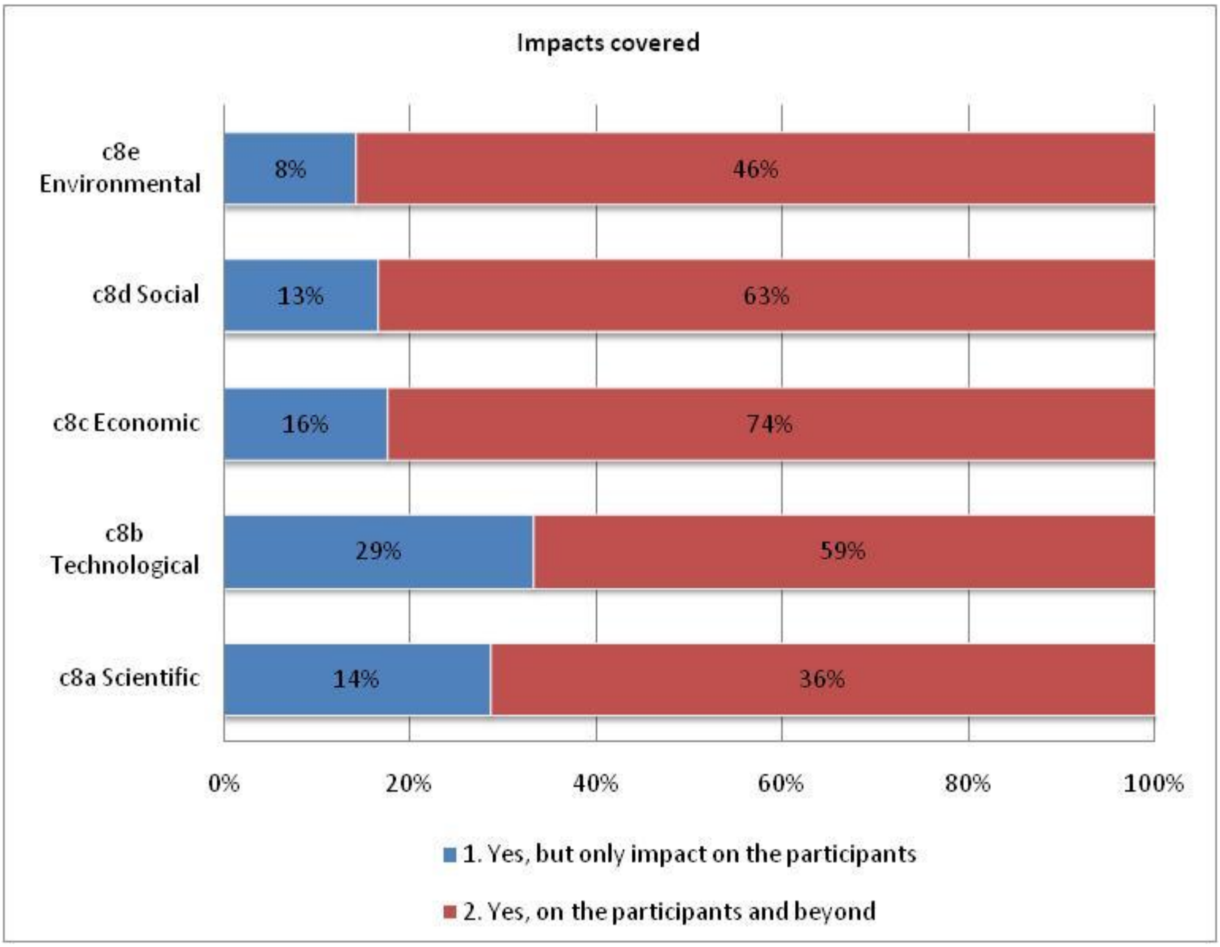
Table 111: Data collection methods covered across the different evaluation types

	Mediterranean countries				Overall results			
Data Collection Methods	Ex ante	Acc	Interim	Ex post	Ex ante	Acc	Interim	Ex post
C.13.a. Existing Surveys/Databases	36,36%	4,55%	22,73%	31,82%	68.18%	70.83%	63.01%	70.83%
C.13.b. Participant Surveys	10,00%	5,00%	25,00%	30,00%	18.18%	66.67%	69.86%	68.75%
C.13.c. Non-participant Surveys	13,33%	-	0,00%	20,00%	18.18%	8.33%	24.66%	22.92%
C.13.d. Interviews	15,79%	5,26%	10,53%	21,05%	54.55%	75%	86.3%	66.67%
C.13.e. Focus Groups/Workshops/Meetings	35,71%	-	14,29%	0,00%	54.55%	50%	43.84%	43.75%
C.13.f. Peer Reviews	7,69%	-	0,00%	0,00%	4.55%	8.33%	19.18%	16.67%
C.13.g. Technometrics / Bibliometrics	0,00%	-	0,00%	8,33%	0%	0%	1.37%	4.17%
C.13.h. Document Search	44,44%	5,56%	27,78%	16,67%	77.27%	41.67%	61.64%	56.25%
C.13.i. Monitoring Data	37,50%	-	25,00%	18,75%	45.45%	91.67%	75.34%	60.42%

2.4 Impacts covered and Main audiences

Most of the impacts recorded are of the technological and the socio-economic type affecting participants, as well as the wider environment. However, it must be noted that according to Structural Funds procedures and forms anticipated, which is the case for most of the evaluations in the Mediterranean group, impacts are defined by indicators, which are very specific to the measure or action. This makes it difficult to classify impacts clearly under 'scientific', 'technological', 'economic', 'social', or 'environmental'.

Exhibit 84: Impacts covered in Mediterranean countries



In terms of the main audiences targeted by the evaluations of the Mediterranean group, external (co) sponsors, auditors and financial authorities, policy analysts and programme management are mainly targeted in ex-ante evaluations; programme managers and financial authorities are involved in both ex-ante and interim evaluations whereas programme managers and government officials are especially engaged in interim evaluations. It is obvious that programme managers have gathered very high shares in all evaluation types within the Mediterranean group.

Comparing to the overall results group, there are no striking differences in the different evaluation types. Ex-ante evaluations do target the same audiences with a minor exception the engagement of 'government officials' in the respective results in the overall group which shows a very high share among all other preferences in this group. The same preference is shown in interim evaluations together with the involvement of 'policy analysts'. A clearer difference is presented in the ex-post evaluations where the overall results group engage not only 'government officials' and 'programme managers' as presented in the Mediterranean group, but also 'politicians', 'those directly supported by the measure' as well as 'policy analysts'. The latter finding suggests that the evaluation discourse in relation to ex-post evaluations is narrower in the Mediterranean countries when compared to the overall results.

Table 112: Intended Audiences covered

	Mediterranean countries				Overall results			
Intended Audiences	Ex ante	Acc	Interim	Ex post	Ex ante	Acc	Interim	Ex post
C.14.a. Policy makers (Politicians)	33,33%	-	27,78%	16,67%	57,90%	62,50%	60,90%	76,90%
C.14.b. Policy makers (Government officials)	33,33%	4,76%	23,81%	28,57%	95,50%	100,00%	98,60%	97,80%
C.14.c. Programme management	40,00%	-	25,00%	30,00%	100,00%	100,00%	97,10%	95,70%
C.14.d. Auditors/ Financial Authorities	47,06%	-	29,41%	17,65%	65,00%	73,70%	45,80%	42,90%
C. 14.e. Those directly supported by the measure	25,00%	-	18,75%	18,75%	40,00%	90,50%	41,90%	57,10%
C.14.f. External/ Internal (co) sponsor of the measure/ programme	53,33%	6,67%	13,33%	20,00%	70,60%	47,40%	40,00%	27,50%
C.14.g. Potential users of the measure	26,67%	-	13,33%	20,00%	29,40%	68,80%	31,70%	42,10%
C.14.h. Policy Analysts	44,44%	5,56%	22,22%	22,22%	72,20%	80,00%	50,00%	53,30%
C.14.i. General Public	16,67%	-	16,67%	11,11%	29,40%	46,20%	27,10%	34,10%

2.5 Quality of evaluations in the Mediterranean countries

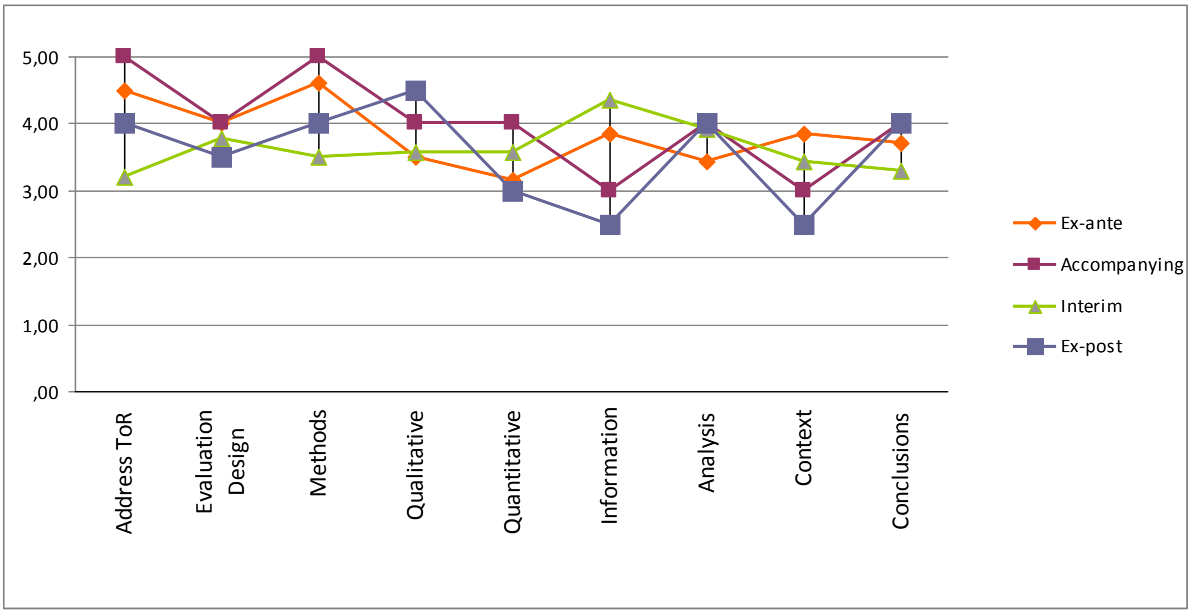
Certain differences appear in terms of quality characteristics across different evaluation types within the Mediterranean group of countries. The high quality traits of **interim** evaluations are comprised of well documented and referenced information sources which are used in the report. The low quality point detected, on the other hand, regards the degree that conclusions are soundly based, as well as the fact that the Terms of Reference (ToR) are actually included in the evaluation.

The high quality points of the **ex-ante** evaluations consist of methods satisfying the purpose of the evaluation (ToR), as well as the fact that the ToR are being addressed within the evaluation. Quality characteristics are lower in terms of analysis as well as application of qualitative and quantitative methods.

The ex-post evaluations indicate that they are in fact better in terms of application of qualitative methods, analysis of findings and providing sound conclusions. They fall short in quality, however, in terms of context coverage and well documented and referenced information sources.

Overall, an encouraging finding is the fact that the quality of almost all characteristics across the evaluation types is graded above the mean (3 on a 5-point scale).

Exhibit 85: Quality of evaluations across evaluation types in Mediterranean group



2.5.1 Distributions across quality categories

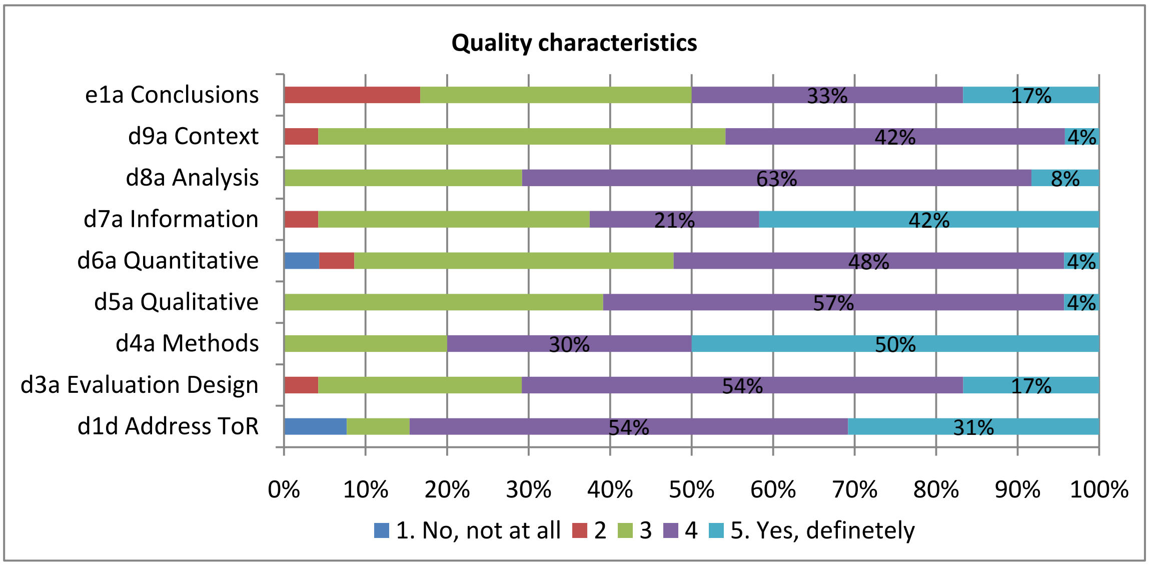
Exhibit 86 shows the distribution for each quality indicator, as far as the Mediterranean group is concerned. They present the perceived quality of the evaluations based on a five-point Likert scale. A high rate of satisfaction can be observed in addressing the ToR within the evaluations, in the fact that methods chosen satisfy the ToR of the evaluations as well as when the analysis is based on the given data. Comparing those results with that of the overall results group the only difference spotted was the fact that the quality characteristic of conclusions based on the analysis occupied a higher share than that in the current group.

Taking only answer 5 (Yes, definitely) into account, the quality characteristics of 'methods chosen that satisfy the purpose of the evaluation' displays the highest share (50%), followed by the 'well documented and referenced information' characteristic with 'address of ToR' coming next. On the other hand, 'analysis based on the data given' comes 3rd in terms of dominance, when adding up answers 4 and 5, but those who have answered 5 in the same category score less than 10%.

The five top quality characteristics in the overall results (presented in section 3) were 'addressing ToR', 'the degree the methods satisfy the ToR', the degree the 'analysis is based on the data given', the degree 'conclusions are based on the analysis' and the appropriateness of the 'evaluation design'. The Mediterranean countries present the same results in terms of the most satisfied quality characteristics with the exception of 'conclusions based on analysis', which is ranked in the last positions of satisfaction.

In terms of satisfaction of each of the quality characteristics, however, all of them are less satisfied in the case of the Mediterranean countries in comparison with the overall results (presented in section 3)

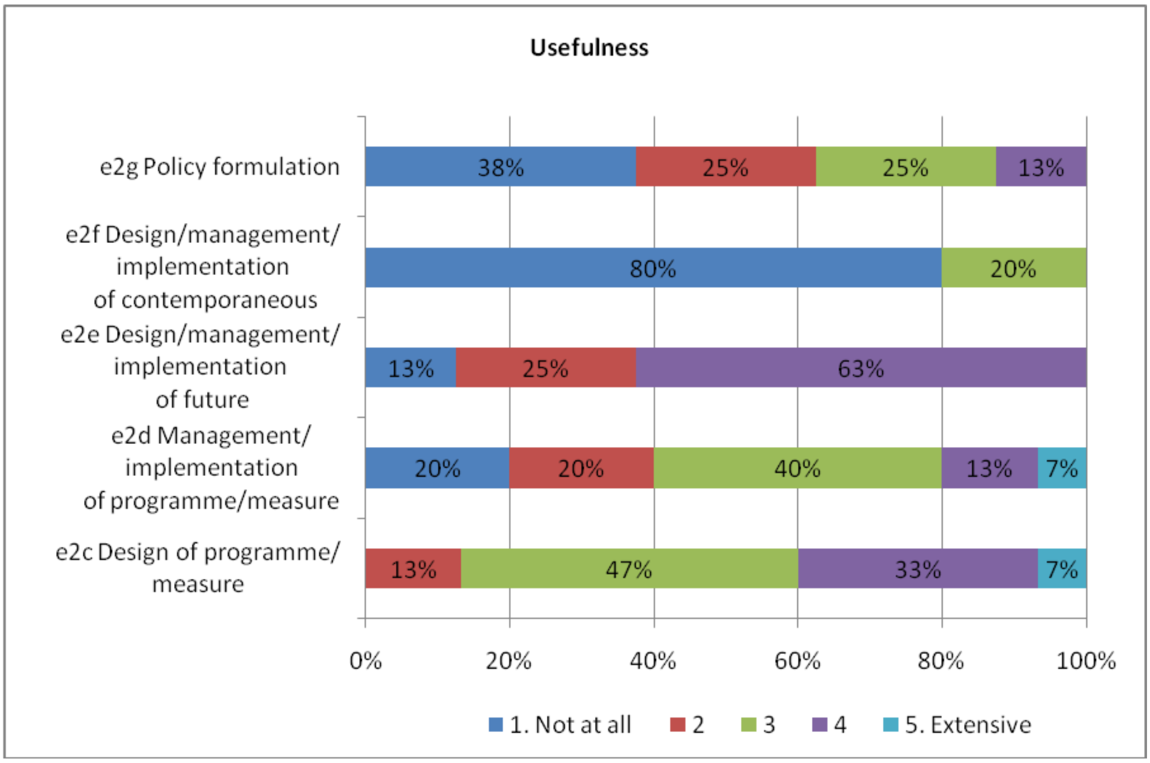
Exhibit 86: Distribution across quality characteristics with the Mediterranean group



2.5.2 Distributions across usefulness categories

In total 18 (75%) of all cases included recommendations. The figure below displays the distribution for each of the "usefulness" categories. It is quite clear that the aspect of usefulness in the Mediterranean countries group mainly relates to the design, management and implementation of future programmes/ measures appraised. The situation in the overall results group is the same, presented in section 3, with design, management and implementation of future programmes/ measures appraised scoring the highest. In addition, changes in the design and management/ implementation of the programme/ measure appraised score relatively higher in the overall results group.

Exhibit 87: Distribution across usefulness categories within the Mediterranean group



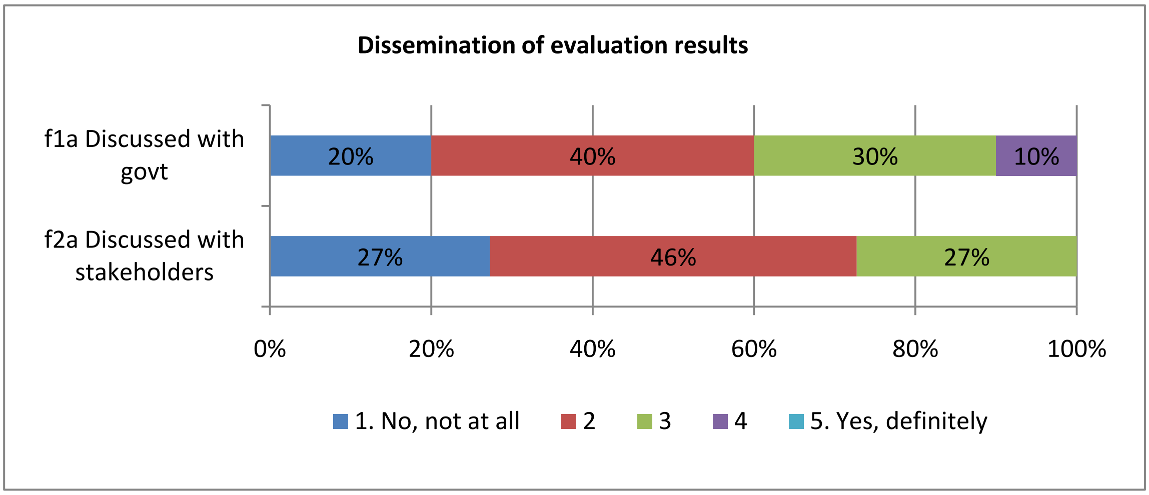
2.5.3 Dissemination

According to a quality criterion set under SF regulations, evaluations should follow an 'Open evaluation process', meaning that relevant stakeholders are to be involved in both the stage of design of the respective evaluation as well as in the analysis of the results. This is an idea which the Mediterranean group of countries does not seem to favor at all

According to Exhibit 88, the results of the evaluations, as far as the Mediterranean group is concerned, seem to be of slight interest in discussions being held within government circles, whereas they are not at all discussed at all among participants and broader stakeholders.

A striking difference appears when the results are compares with the overall results (presented in section 3). Those revealed that in around 50% of the cases results were discussed with government officials and the same share referred to discussion with wider stakeholders (scoring 4 and 5 on an 1-5 point scale). In the Mediterranean group as noted above the respective shares 10% and 0% respectively.

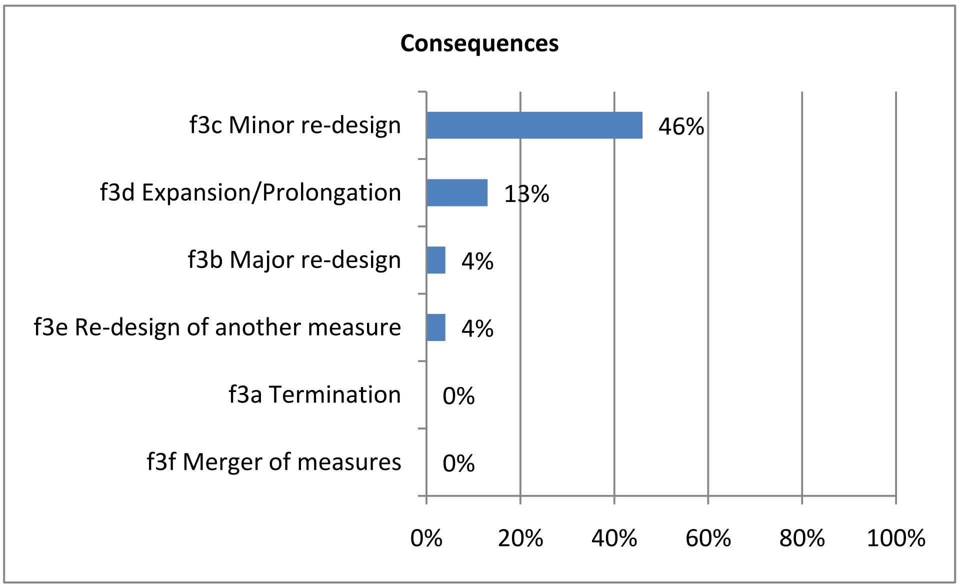
Exhibit 88: Dissemination of evaluation results



2.5.4 Consequences

Exhibit 89 shows the distribution for each of the anticipated type of consequences that the different evaluations appoint to within the Mediterranean group. The most frequent consequences observed are mainly the ‘minor re-design’ of the measure with a 46% share followed by the ‘expansion/ prolongation’ of the measure presenting a lower share of 14%. Comparing those results with that of the overall group, (section 3) no differences are spotted. ‘Minor re-design’ of the measure as well as its ‘expansion/ prolongation’ are the most frequent consequences detected, with the rest being very seldom.

Exhibit 89: Consequences of evaluations in the Mediterranean group[179]



---

[179] Percentage of “Yes” response in Sample B.

3 Conclusions

The survey has indicated a small number of differences, but mainly across the different evaluation types, rather than across the Mediterranean countries and the overall population. More specifically, the main conclusion, drawn upon studying the traits of different types of evaluations (ex-ante, interim and ex-post), is that there are in fact distinct differences in terms of evaluation topics and that the situation is more uniform with regards to data collection methods and analysis, as there are slight differences among evaluation types. The situation is rather similar in the overall results.

Regarding the Mediterranean countries' group, which form the focus of the case study, these differences can be attributed to Structural Funds regulations. However, the fact that the same differences across the different evaluation types are also spotted in the overall results suggests that SF regulations do guide the way evaluation types are conducted; yet they do not prescribe something different from what is usually adopted in international practice.

As far as evaluation audiences are concerned, there seems to be a variety of audiences corresponding to each evaluation type, this being the case for both groups. The situation was also similar in the case of which recommendations were considered more useful or the type of the most usual consequences of the evaluations.

In terms of quality, the adoption of SF regulations has had a positive impact in these countries where there is lack of evaluation tradition. This is based on the fact that almost all quality criteria score between 3 and 4 on a 1-5 point scale in terms of satisfaction. However, these scores are lower than the overall results.

Despite the relatively good quality of evaluation in the Mediterranean group the results are rarely discussed with government cycles or broader stakeholders. This forms a striking difference with the overall results where almost half of the cases reported discussion of results with government cycles or broader stakeholders.

This page has been intentionally left blank.

Part IV
Chapter 12

Conclusions and Ways Forward

Jakob Edler and Ken Guy[180]

---

[180] The lead authors for this section of the report are as given above, with contributions from all other team members at different parts of its production, conclusions and interpretations have been approved and are shared by the group.

1 Towards Lasting Value: INNO-Appraisal and its Repository in Historical Context

Policy and programme evaluation in the field of R&D and innovation has come a long way in the last quarter of a century. Or has it? Some things have obviously changed enormously, which seems to make the query redundant, but what does closer inspection reveal?

At the beginning of the 1980s, the terms policy evaluation and programme evaluation were rarely used in the R&D and innovation sphere in Europe. Indeed, funding for R&D and innovation was only just starting to be organised via programmatic interventions, and although the concept of policy evaluation was common in fields such as education and health, especially in the USA, neither the concept nor the practice of policy evaluation were much in evidence in Europe. The terms 'evaluation' and 'assessment' were used, but mainly in terms of assessing the performance of scientific fields (e.g. via bibliometric approaches), or technological developments (via studies of patenting in different sectors and countries) or even evaluations of the performance of individual institutions (e.g. scientific laboratories). But very few studies focused specifically or systematically on the evaluation of government policies in the R&D and innovation realm.

Things started to change, however, with the spread of an R&D and innovation programme culture in the early 1980s (e.g. the Alvey Programme for Information Technology in the UK and the first EU Framework Programmes such as the ESPRIT programme, again in Information Technology). And concurrent with the launch of initiatives such as these came a political demand for increased accountability and greater return on public investment in R&D and innovation which in turn led to a demand for evaluations of these initiatives and a search for people capable of conducting them.

One of the first places that public officials turned to was the academic sector, particularly to the few centres that had started to specialise in 'science policy' and 'science in society' issues in the 1960s. Even there, however, there was little expertise in 'policy evaluation' and none in 'programme evaluation', since programmes themselves were still something of a novelty in Europe. There was enthusiasm, however, and a cadre of young researchers excited by the prospect of being in at the start of a relatively new phenomenon.

And exciting times they were. Some of the most important steps to be taken involved the development of the language of evaluation and the conceptual frameworks linking terms such as efficiency, effectiveness, relevance and appropriateness; inputs, outputs, outcomes and impacts; input, output and behavioural additionality etc. All these terms are now commonplace in the world of R&D and innovation evaluation, but only because of the early efforts to systematise ways of thinking about the theory and practice of evaluation.

As the language and concepts of evaluation evolved, so too did the methodologies deployed. The basic triumvirate of literature review (programme documentation, policy documents, academic literature); surveys (mainly of project participants); and interviews (with programme administrators and participants) were complemented with other approaches and levels of sophistication rose (especially in terms of questionnaire design and analysis; the use of network analysis tools to explore the spatial and temporal development of collaboration patterns; the use of econometric techniques to analyse downstream impacts on participants and the broader environment; the use of cohort techniques to track the comparative performance of successful, unsuccessful and non-applicants etc.).

But there have also been many other changes, especially in terms of the growth of an evaluation culture amongst both the ‘policy analyst’ community responsible for performing evaluations (primarily academics and consultants specialising in R&D and innovation and/or evaluation issues) and the ‘policy maker’ community responsible for commissioning evaluations. Both communities were thin on the ground and fairly amateurish in the early days (as far as evaluation was concerned), but now both communities are larger and more professional. Nowadays, for example, some invitations to tender for evaluations reflect a sophisticated understanding of evaluation issues and techniques. Equally, some of the responses by the policy analyst community continue to exceed expectations in terms of the novelty and variety of the techniques they are interested in developing and applying.

On the face of it, therefore, much has changed for the better. But a more detailed look at the policy analyst and policy maker populations and their respective degrees of sophistication reveals a less satisfying picture. Certainly vanguard developments by the analyst community are often impressive, but elsewhere there is evidence of evaluations being performed mechanistically, with no real insight or understanding of the subject matter being evaluated; of a lack of consensus over the meaning of basic evaluation concepts such as efficiency and effectiveness; and even evidence of a certain degree of ‘reinvention of the wheel’, which reflects an unfortunate, albeit enthusiastic, perpetuation of amateurism. Similarly, many invitations to tender for evaluations continue to reflect a weak understanding of what is, and what is not, possible to achieve within the scope of a typical evaluation.

To some extent this is not surprising. Both communities are growing as the need for evaluation becomes an embedded part of the strategic intelligence function of governance systems around the world, but it is not obvious that the mechanisms in place to train either policy analysts or policy makers in the increasingly ‘less-black’ arts of evaluation are fit for purpose. Professional training courses exist, but few amongst the analyst and policy maker communities have taken advantage of them. The number of people holding an ‘evaluation brief’ for more than a few years during a working career is also relatively small, and mechanisms designed to preserve ‘institutional memories’ are frequently weak. The picture that emerges when looking at both the policy analyst and policy maker communities, therefore, is one of a few ‘hot spots’ of evaluation excellence and long tails of relatively inexperienced practitioners lacking adequate reference points. Complicating the picture even more, the need for learning amongst evaluators and commissioners of evaluations increases daily as the demands on innovation policy to deliver grow and efforts to identify impacts intensify.

Rectifying this growing gap between experienced policy makers and evaluators on the one hand and the growing number of inexperienced policy makers and evaluators on the other is where INNO-Appraisal can be of most value; where it can make the most difference. To date, there has been no authoritative source that either of these growing populations could turn to that documents and codifies practices in the way that the INNO-Appraisal repository now does. Certainly there are guidebooks and manuals that describe evaluation concepts, methodologies and analytical techniques, and there is now an appreciable academic literature on evaluation, but the most numerous and useful sources of information – namely evaluation reports themselves – have to date been firmly embedded (some would say buried) in the relatively inaccessible 'grey literature'.

Though not any more, even the most inexperienced commissioners of evaluations can now read and compare evaluations of programmes similar to their own 'at the touch of a button'. Likewise, policy analysts new to the field of evaluation can easily access the work of established evaluators. Moreover, INNO-Appraisal also allows both analysts and policy makers to develop a range of useful overviews. Commissioners of evaluations, for example, can quickly develop an overview of the evaluation topics covered in different types of programmes and the results one can expect as a consequence. Policy analysts, too, can quickly see what methodologies have been used to address different evaluation issues.

In short, INNO-Appraisal codifies much of the tacit knowledge that currently exists about evaluation practices and acts as a repository for this knowledge. It thus constitutes a source of learning for newcomers, a reference point for experienced practitioners and one way of helping to overcome problems associated with porous institutional memories.

Useful as it promises to be, however, a word of warning needs to be sounded. INNO-Appraisal will only be useful today if adequate steps are taken to publicise the existence of the repository and its utility as an analytical tool to all members of the policy analyst and policy making communities interested in evaluations. Perhaps more importantly, however, INNO-Appraisal will only be useful tomorrow if an adequate maintenance and up-dating strategy is developed. If this is not done, the utility of INNO-Appraisal is likely to decline exponentially with a very short half-life, and all the effort and resources devoted to its construction will have been wasted. The INNO-Appraisal team strongly recommends that efforts are made to ensure the survival of an institutionalised learning tool for evaluation and innovation policy in Europe. INNO-Appraisal should also be seen as a starting point for greater self-reflection by the evaluation community, with many more in-depth studies needed on evaluation practice and its contextualisation.

2 Developments in evaluation practice

This study has, for the first time, provided the policy community and the evaluation community in Europe with a statistical account and analysis of evaluation practice in Europe. Evaluation practice in Europe is highly diverse: it differs between countries and it shows an enormous range in terms of methodological approaches, coverage of topics, quality and usefulness. Different institutional settings and policy traditions in countries influence evaluation practice – and vice-versa, as especially the Austrian case has shown. Evaluation has spread across Europe as the structural fund provisions have pushed countries towards evaluation – though with mixed results to date. The analysis presented in this report constitutes an important step forward in our understanding of evaluation. One key consequence, or so the authors of the study hope, is that the results will allow both policy makers and evaluators to reflect about their own practice, about their approach to evaluation and, ultimately, about the use of evaluation.

While readers may draw their own conclusions as to the lessons to be learned from the analyses presented in this report, and while each of the chapters delivers specific insights from which lessons can be drawn, there are a set of key observations that should support further improvements in evaluation practice across Europe. Once a rarity, evaluations are becoming increasingly commonplace, yet the analysis has shown that this does not automatically lead to good quality evaluations and productive learning as a consequence of evaluations. Greater care needs to be taken along the whole policy cycle to ensure that evaluations are correctly designed, implemented and utilised, with close interaction at all stages between those commissioning and those performing the evaluations. Policy makers need to be 'intelligent costumers', they need to have the absorptive capacity to understand what evaluations can deliver and what they cannot deliver. Evaluators, in turn, must ensure quality throughout the process, especially, though not exclusively, in the application of methods and the development of a thorough understanding of the wider policy and political context in which measures are situated.

Further, conditions and practices concerning the discussion of evaluations within government and beyond must be improved. More thought needs to be given at the planning stage to this phase of the process and to the channels of communication that can be exploited, but evaluators themselves also have to bear in mind that the likelihood and quality of subsequent discussions are highly dependent upon the perceived quality of their reports and the clarity with which methodologies are described and results presented. All this then leads to a more fruitful discussion within and across government and better-informed decisions. In future, however, there will be a need for even greater conceptual clarity given the increasing complexity and sophistication of both innovation policy and the evaluation tools needed to assess the impacts of these developments. The case study of behavioural additionality demonstrated how complex it is to turn one important idea into an operational concept that is both theoretically sound and offers added value to policy makers.

Other operational improvements are also needed. These include the more tailored and conscious design and use of monitoring systems, with evaluations building on the data they produce and monitoring becoming an integral part of the learning process. Evaluation, moreover, should be perceived as a mobilising tool for innovation policy at large, , a function highly underused.

Finally, a dilemma confronting evaluation has to be noted. In order to provide the new methods and concepts needed to better inform policy, evaluation itself has to be innovative. Yet the commissioners of evaluations are often very conservative, specifying conventional methodological approaches in their terms of reference despite known limitations and shying away from more experimental approaches. Opportunities to push the boundaries of evaluation theory and practice are thus often constrained.

Allowing for more experimentation, however, will become more important in the future. Evaluation practice in Europe will have to follow the principle of 'form follows function' much more closely. The evaluation of innovation policy will have to adapt to new trends in innovation policy and the demands being placed upon it. The analyses in this report have shown a considerable degree of uniformity of evaluation designs across policy measures. Evaluation practice, to a large degree, is an exercise in 'copy and paste' into new application areas. However, policy measures are likely to differ even more in the future, and evaluation will have to adapt. To highlight one key example, one major trend is the increasing importance of demand-driven innovation policy and diffusion-oriented measures. For these, evaluation practice is almost non-existent. This has a set of implications. Evaluation will have to tackle systematically and with methodological rigour a broader range of impacts – the focus on technological and economic impacts is increasingly too limited. Our understanding of how demand-side drivers and policies can interact with and influence supply-side developments also needs to improve radically before adequate evaluation approaches can be developed, and this understanding has to be shared by policy makers and evaluators alike.

A second example concerns the vastly increased emphasis the structural funds place on innovation, where there is a clear need for new innovation concepts in extremely challenging environments.[181] Without the development of intelligent and appropriate evaluation concepts and practices along the policy cycle, there is the danger that new application areas and innovation policy instruments might be supported by evaluation practices that are transferred without any consideration for contextual differences or – even worse – driven by ideological preconceptions. Hopefully, however, the lessons from INNO-Appraisal, the discourse we hope to support and the learning tool we provide can be of some assistance when designing and implementing improved and tailored evaluation approaches that will be needed in the future.

---

[181] First discussions between the INNO-Appraisal team and officials from DG Regio were held on February 4 2010 concerning the transfer and further development of concepts for structural fund evaluations in the area of innovation policy.