

The practice of evaluation in innovation policy in Europe

Jakob Edler^{1,*}, Martin Berger², Michael Dinges² and Abdullah Gök¹

¹Manchester Institute of Innovation Research, MBS, University of Manchester, Manchester, UK
and ²Joanneum Research, Vienna, Austria

*Corresponding author. Email: jakob.edler@mbs.ac.uk

This article characterizes and analyses evaluation practice in national innovation policy across Europe. It is the first study that examines and interprets the characteristics, quality, usefulness, and consequences of evaluations in a systematic way. The analysis is based on the comprehensive INNO-Appraisal repository of 171 evaluation reports of national innovation policies of EU25 countries, conducted between 2002 and 2007. The article seeks (1) to assess the state of the art of evaluation in innovation policy at national level, (2) to understand how different key dimensions of evaluation (timing, purpose, methods, tendering process, etc.) relate to each other, and (3) to explore types of evaluations. On that basis, we (4) draw lessons as to what constitutes good practice in evaluation, as the results of the survey have been exchanged and discussed with a number of policy makers of the sample responsible for the evaluation. The article thus both contributes to the academic understanding of policy evaluation and supports use in policy practice.

Keywords: evaluation; innovation policy; evaluation practice; evaluation typology.

1. Introduction

Innovation policy has entered centre stage of public policy in Europe, and it has become enormously diversified across Europe. The number of measures at national level has grown over the past decades, as more and more interventions have sought to tackle different aspects of perceived market and system failures.¹ In this article, we follow the scope of the EU Trendchart database and define innovation policy instruments as all systematic measures of public policy actors that are geared towards the improvement of capabilities to generate and apply new products, processes, and services. Innovation policy in this understanding is primarily targeted towards firms, but includes schemes that link public science organizations with firms or that tries to exploit knowledge generated in public research. Our (narrow) understanding does not include science policy, which is geared towards the generation of scientific knowledge mainly, thus the database does not include schemes that exclusively target public research organizations that produce academic knowledge.

The objectives and intervention mechanisms of innovation policy have broadened in scope. 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

(Arnold 2004; Edler et al. 2008; Flanagan et al. 2011). Intervention rationales in innovation policy are based on a set of theoretical assumptions as to what drives innovation capabilities and performance, and how improved capabilities and performance lead to technological, environmental, social, and economic impacts. This illustrates the high expectations regarding the effectiveness of innovation support measures. In addition, constrained budgets make for more pressing choices between interventions: any intervention needs to be implemented as efficiently as possible. Against this background, evaluation has become increasingly important, both as a policy and management supporting tool and as a tool to assess policies in order to justify or re-direct funding. At the same time, however, we have a very fuzzy and incomplete understanding to what extent, and in what form, the ever-growing aspirations of innovation policy are supported by appropriate analytical and formative means.

This article is the first systematic analysis of the state of the art of evaluation practice in innovation policy across Europe. Its starting point is that although we know the demands for evaluations (Miles et al. 2005), we do not have a clear picture at all about the overall evaluation practice. This article intends to fill this gap by building on the results of a longer term project to take stock of

and analyse evaluation practice in innovation policy in EU countries (Edler et al. 2010). It develops a phenomenology of evaluation practice and offers the first step towards a typology which is informed by how key characteristics of evaluations are linked to each other.

Our study on the evaluation practice in Europe is to be seen in a historical perspective. It builds on earlier attempts to understand needs and practices of the evaluation of innovation and technology policy in a comparative fashion, but follows a different methodological approach and can draw on more systematic data. One of the earlier key contributions was the OECD conference on policy evaluation in innovation and technology 1997 (OECD 1999), as it developed a structure to understand and compare evaluation (Papaconstantinou and Polt 1999) and provided a whole range of country practice analyses. This exercise demonstrated already a trend towards broad evaluation approaches that were in-built in many programmes (Georghiou 1999b). Since the 1990s, one can observe some convergence regarding the needs for evaluations and what can be regarded as good practice. This was reinforced by further attempts to capture the nature of evaluation, often in the form of comparisons of different country approaches or types of evaluations (e.g. Shapira and Kuhlmann 2003; OECD 1999). The growing relevance and sophistication of evaluation has led to a number of handbooks for evaluation practitioners in science and innovation (e.g. OECD 1998; Fahrenkrog et al. 2002; Ruegg and Feller 2003; Miles et al. 2005).

Most of the preceding exercises to discuss evaluation practice in innovation policy have focused on good practice or delivered country analysis in a more qualitative manner. They did not and could not draw on comparable data across a set of countries to ascertain and analyse the actual state of the art of evaluation. This article does just that. It uses techniques of Meta-Analysis² 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). It does *not* systematically gather and synthesize information from evaluations in order to better understand the policy measures that are evaluated (as mostly done in meta-evaluations described in Georghiou 1999a and Edler et al. 2008), but to better understand evaluation practice.

This analysis focuses on evaluations of innovation policy as conceptualized within the European database on innovation policy EU INNO-Policy TrendChart.³ Innovation policy is thus understood to comprise public action that tries to enhance the innovation capabilities and performance of private and public actors, both by targeting those actors directly, and by setting up intermediaries and framework conditions that benefit the target groups directly. Thus, the analysis does not include evaluations on science policy instruments or organizational evaluation.

The article starts off by developing a phenomenology which defines the major categories used to characterize evaluations and by explaining the nature of the data and the methodology (Section 2). On that basis it presents the data on evaluation practice in innovation policy by way of descriptive statistics (Section 3). Two further sections (Sections 4 and 5) analyse key dimensions of evaluation in more depth. The first one discusses how the methods employed are linked to the topics and impact the dimensions they cover (Section 4.1) and how evaluation designs differ for different kinds of policy measures (Section 4.2) as well as different kinds of purposes (formative versus summative). Section 5 explains what determines the perception of quality of evaluations by policy makers (Section 5.1) and, subsequently the consequences that arise from evaluation for policy design (Section 5.2). A final analytical Section 6 presents the results of a cluster analysis. This reveals three types of evaluations: the ‘verdict’—a summative evaluation with limited breadth and usefulness; the purely ‘supportive’ approach that focuses on programme implementation and qualitative methods, and the ‘holistic’ approach that combines assessment with formative purposes. These three types are then analysed in more depth. A final section summarises and interprets the major findings and recommends ways forward to improve evaluation practice in innovation policy.

2. Capturing evaluation practice

2.1 Categories to characterize evaluations

Evaluations are unique, idiosyncratic exercises. Each evaluation has its specific political, stakeholder, and policy context, with specific requirements defined by those commissioning it and depending on its role in the policy cycle. The goals of evaluations differ and with this the dimensions that are covered and the methods that are used. However, if we want to understand evaluations beyond the idiosyncratic case, and on that basis capture and analyse evaluation practice, we need to define a limited set of variables that can be used to characterize evaluations and to analyse how different aspects of evaluations are linked.

Hence, for the underlying study a specific data capture concept has been developed. Figure 1 summarizes the principle dimensions that were used to characterize evaluations. *Policy Measure* is characterized in terms of objectives and target groups. The characterization of the policy measure is used to test whether distinct policy measures require/trigger certain evaluative questions and methods. *Evaluation Set-up* characterizes timing, purpose, and commissioner of the evaluation. Here, the relation between different characteristics and main questions and methods can be analysed. *Main Questions* provide a categorization of evaluative topics and impact channels covered by an evaluation. *Methods* provide a

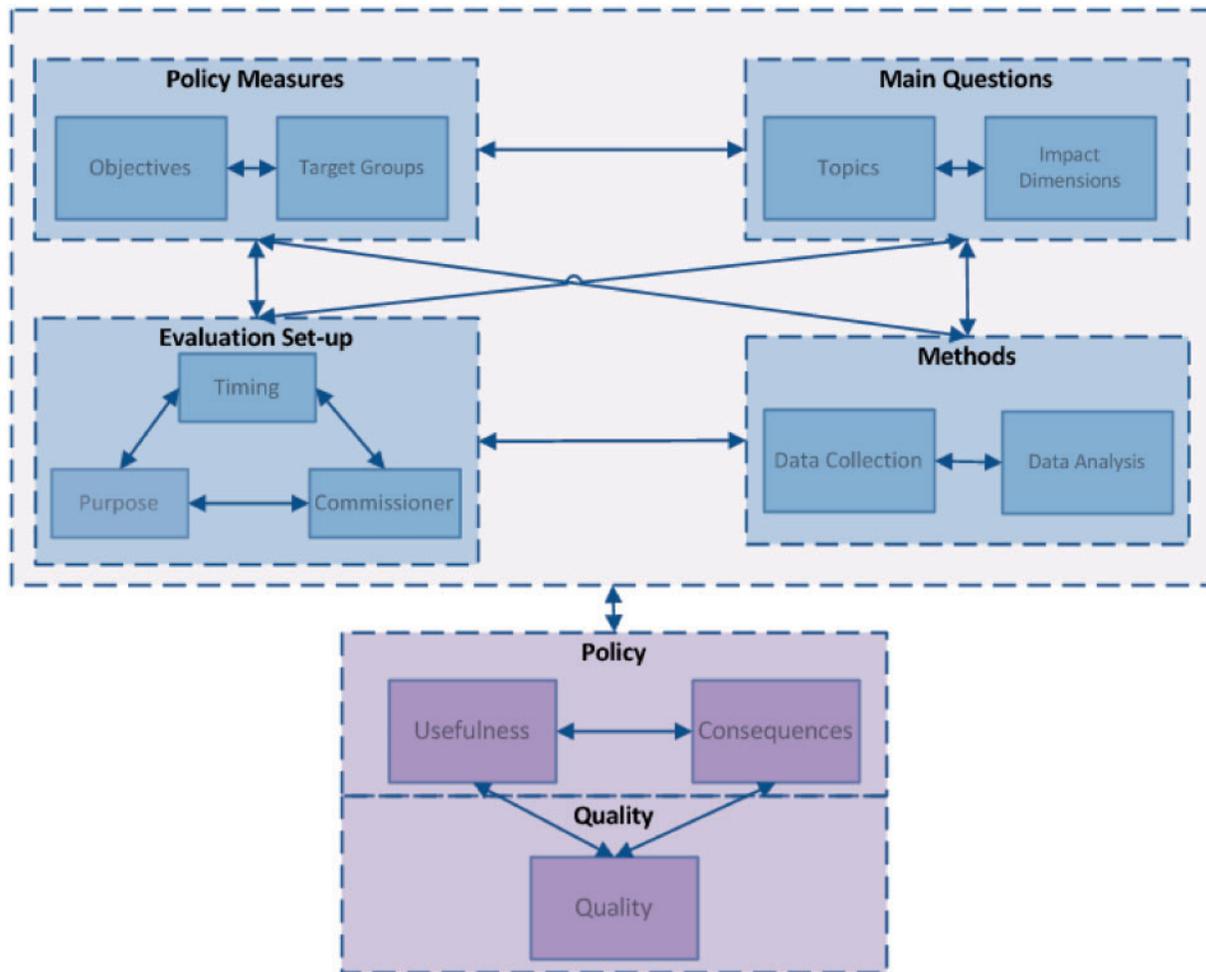


Figure 1. Key evaluation dimensions and their relations.

categorization of data collection methods and data analysis approaches used in an evaluation. *Quality* is a dimension that is subjective, determined by the authors together with the policy makers. All these categories are linked to the *Policy Dimension* which provides information on the usefulness and consequences of the corresponding evaluations.

2.2 The data and methodology

The INNO-Appraisal database contains evaluations of a whole range of different innovation policy measures that are covered in the European INNO-Policy TrendChart Database between 2002 and 2007.⁴ Figure 2 depicts the breadth of underlying innovation policy measures that are represented in the database. The typology of measures largely follows the one developed in the INNO-Policy TrendChart.

The INNO-Appraisal database covers innovation policy instruments in EU25 countries⁵ but not, at the EU level. Although the database is the most comprehensive database

on evaluations carried out in the European Union Member States, it cannot claim to cover all innovation policy evaluations in all countries to the same degree, since the coverage of policies in those countries is not equally complete for all countries, reporting practices in the database differ for countries. The INNO-Appraisal database also extensively covers evaluations of structural fund measures: slightly more than 20% of all evaluations in the repository are performed in the context of structural funds.⁶

The basis for the characterization of the evaluation reports consists of all reports and related documents that could be gathered by the study team and a network of correspondents. This exercise was conducted not only through a search of publicly available sources but also via communication with relevant INNO-Policy TrendChart correspondents and in some cases the respective policy makers. For each evaluation, a characterization template covering the variables discussed above was filled in by the study team and subsequently checked and complemented by the policy maker responsible for that evaluation report.

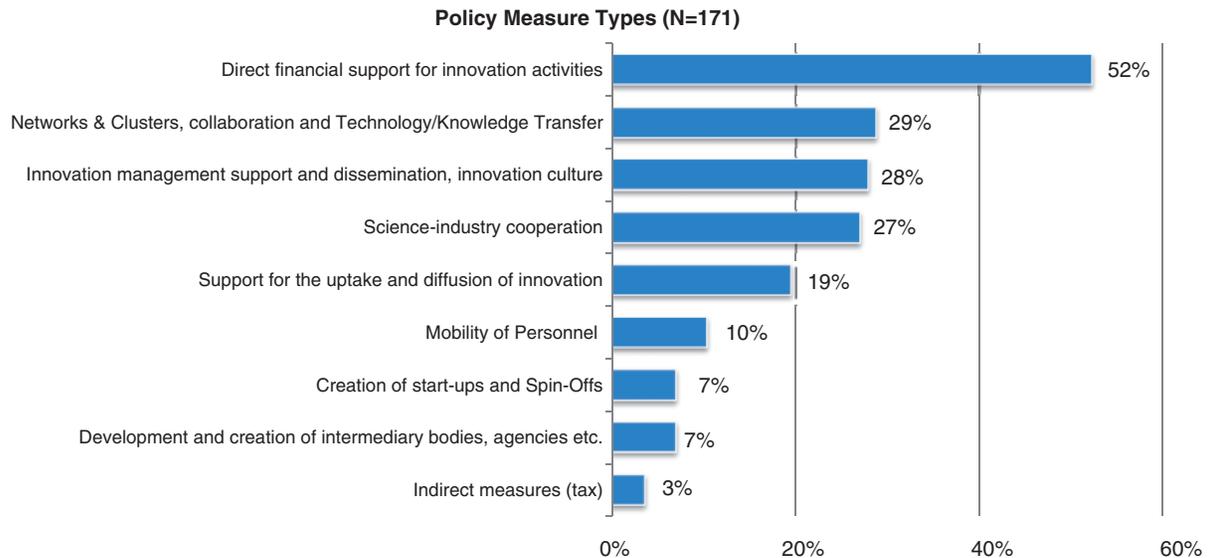


Figure 2. Types of policy measures represented in the INNO-Appraisal database (share of policy measure type represented in evaluations in dataset, multiple allocations possible).

The policy makers also added information that was not evident from the report or other sources and gave their own, personal account of the quality, the consequences of the evaluation, and the usefulness of the recommendations.⁷

This led to 242 characterized evaluation templates in the INNO-Appraisal database, 171 of which could be completed with sufficient quality and coverage of variables. Of those 146 were again amended and verified by policy makers and 132 of those, in turn, were of sufficient detail and quality to be used for the analysis. Consequently, for some statistical analysis we will use the larger set which contains all the datasets, including those that contained only entries by the study team (171 cases). For the questions on quality, usefulness, and consequences we will turn to the smaller dataset of 132 cases. Methodologically, we analyse the association between two variables either using cross-tabulation or correlation analysis.

3. The big picture: overview of evaluation practice

3.1 The nature of evaluation practice

The overall dataset allows exploring main features of evaluation practice in Europe. Evaluation is found to be an integral part of innovation policy in our sample of innovation policy in European countries. Roughly 50% of the measures that are evaluated have a pre-determined budget for evaluation and two-thirds of the evaluations are foreseen and planned during the design phase of the measure.

The close ties between programme design and evaluations are also reflected in terms of contractors. More

than 90% of evaluations are sponsored by the programme owners themselves, only a minority are jointly sponsored with other bodies or entirely externally funded (10%). In those cases, most often co-funding takes place via the European Commission through structural funds (see Amanatidou and Garefi 2011). Also, the intended audience of evaluations points towards this direction: government officials (98%) and programme management (98%) directly in charge of the support programme constitute the main intended audience for evaluations. Interestingly, those directly supported by the measure and potential users of the measure are only targeted in about half of the evaluations, which means that the potential to mobilize the community does not appear to be fully exploited.

In terms of commissioning evaluations, the analysis reveals that evaluations are by and large external services which are procured by the respective authorities via tender procedures and clearly specified objectives. ‘External’ here means that the evaluations are performed by consultancies, academics, or other administrations that are independent from the organization that has designed or implemented the measure. Almost half of the evaluations followed an open tender procedure and roughly 20% went through a closed tender procedure (with a limited number of invited tenderers). Another fifth are performed by external evaluators without a tender procedure and 15% of the evaluations were carried out internally. In general, tendered evaluations had clearly specified objectives, while at the same time two-thirds of the tender documents left the choice of methods to the evaluators.

More than 40% of all evaluations are interim evaluations and another 30% ex post evaluations. The bias against ex post (30%) may be partly attributed to the

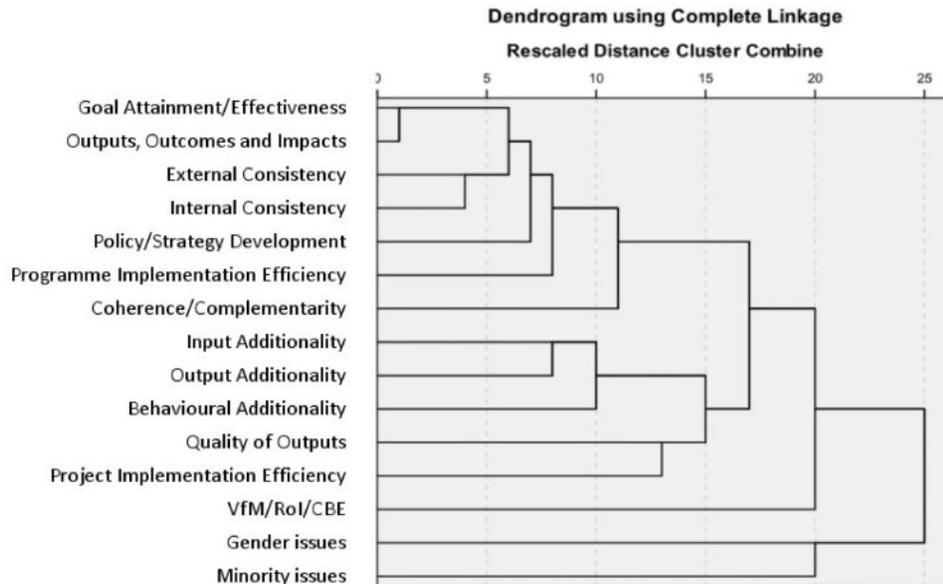


Figure 3. Dendrogram of topics covered in the evaluation reports analysed (average linkage between groups). Řezanková (2009) recommends ‘Jaccard’s co-efficient’ or ‘Yule’s Q’ measures for object clustering (clustering of variables of same type) of dichotomous (variables that take binary options) asymmetric (‘1’ and ‘0’ values are of inherently different importance) variables. This method does not cluster variables on the basis of the co-absence of the same trait (i.e. both variables take the value ‘0’ at the same time). In this analysis, a furthest neighbour method which links topics with complete linkage is used by applying Jaccard’s co-efficient measure. Note: the order of topics from top down also reflects *roughly* the order according to frequency of occurrence.

fact that the underlying INNO-Policy TrendChart database focuses on ‘live’ policies within a certain period of time. Based on expert judgement and policy-makers confirmation, the majority of evaluations tend to combine formative (e.g. learning and improvement oriented) and summative (judgemental) aspects. Roughly 30% of the evaluations are formative only and one-fifth purely summative. We come back to this crucial distinction in Section 4.

As regards topics that are covered in evaluations, our analysis can define their relative importance and how topics link to each other. Around 90% of evaluations analyse ‘outputs, outcomes, and impacts’ and ‘goal attainment and effectiveness’. Those topics are highly related, if one of the two topics is covered, the other one is highly likely to be covered as well (Figure 3). A second cluster of topics is ‘internal’ and ‘external consistency’, which both are included in 80% of the evaluations. As our dendrogram below shows (Figure 3) the most important four topics form a large cluster. Further, they are strongly linked to the topic ‘policy/strategy development’ (76% of evaluations), ‘programme implementation efficiency’ (76%), and ‘coherence/complementarity’ (72%). About two-thirds of all evaluations cover at least one form of additionality (input, output, and behavioural), and those three types of additionality are often looked at together in evaluations. Further, evaluations looking at additionality tend to include the level of individual projects, which can be explained by the need to look at the behaviour of individuals in order to analyse behavioural additionality,

i.e. the changes in behaviour that can be attributed to the policy intervention. ‘Gender issues’ (24%) and ‘minority issues’ (7%) are least common.

Our survey asked about the coverage of four distinct impacts (technological, economical, social, and environmental),⁸ for a range of pre-defined topics and it also asked if the evaluations looked at impact beyond the programme participants. Technological and economic impacts are most often reported as being important, and environmental impacts are least frequently mentioned as important. Interestingly, across all impact dimensions the share of evaluations which claim to look beyond the project participants is higher than those that are limited to the participants only. This appears to reflect the growing need to demonstrate the societal and broader economic benefits of policies.

The evaluation of innovation policy across Europe uses a wide range of methods. In terms of data analysis methods, advanced quantitative approaches such as control group approaches (20%), counter-factual approaches (22%), cost/benefit approaches (23%), econometric analyses (23%), and input/output analyses (26%) are limited in use. On the other hand, simple descriptive statistics (76%) as well as qualitative approaches of context analyses (67%), document analyses (52%) and case study analyses dominate the data analysis methods employed.

As regards the data collection methods employed, 80% claim to use monitoring data and 70% to use existing surveys and databases as a basis for the analysis.

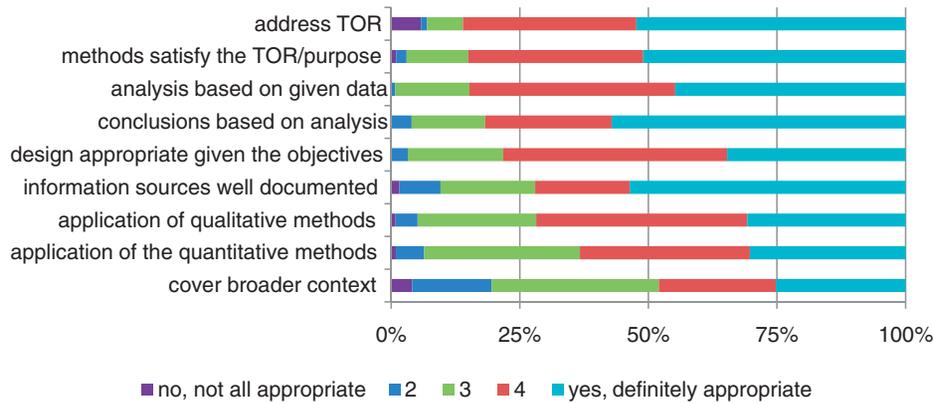


Figure 4. Appropriateness of evaluations: policy makers' assessments. Note: sorted in descending order for the top two categories.

However, it appears that this kind of data is insufficient for specific evaluation questions, such as networking or behavioural additionality. The other data collection methods that were most often employed are interviews and participant surveys. Despite claiming to look at impact dimensions going beyond the participants of evaluations, non-participant surveys are only performed in 25% of all evaluations. Bibliometric and technometric searches are almost completely absent from the observed dataset (2%), while 20% of evaluations use peers to support the assessment of technological substance and management plausibility in projects (20%).

3.2 Quality and consequences: policy makers assessments

The INNO-Appraisal database allows insights into the quality of evaluations. This was done in a subjective way, assessing quality according to eight criteria through analysis by the team and having it confirmed or amended by the responsible policy maker. Thus, while this approach cannot deliver a fully objective assessment following clearly specified criteria, it does help to understand how policy makers have perceived the notion of quality.

The reported distribution of the eight criteria used to define quality is shown in Figure 4: it presents 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 were less satisfied with the coverage of the analysis of the broader contexts (e.g. societal, institutional, policy, and economic contexts) given the objectives of the evaluation, the application of quantitative and qualitative methods, and the documentation of information sources.

In order to operationalize quality perception, a simple binary quality index has been constructed. Out of the eight quality variables, we focused on three variables that sufficiently represent overall quality, namely appropriate design, analysis clearly based on given data, and

conclusions based on analysis. All evaluations that score more than 3 on a Likert scale (1 being very low, 5 being very high) in each of the three selected quality variables are defined as being of high quality. Sixty-one percent 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 dimension. This finding is confirmed through a correlation analysis: many evaluations are either good in a whole set of quality criteria or perform rather badly across the board. The determinants of quality will be explored in Section 5.1 below.

Most evaluations contain recommendations for policy and programme management, only a minority of evaluations are purely analytical. The usefulness of these recommendations was again assessed by the policy makers themselves, based on a set of the following five dimensions: (1) changes to the design, (2) changes to the management and implementation, (3) changes to future programmes, (4) changes to other contemporaneous programmes, and (5) changes to broader policy formulation. While recommendations for management and implementation issues (both for the evaluated programme and future programmes) had high scores, considerable lower scores were tracked for changes in the design of other programmes and broader policy formulation. This indicates a limited spill over of learning to more general policy issues.

The data also allows linking evaluations to consequences, i.e. policy makers were asked if the recommendations made in the evaluation had led to consequences for the evaluated programme or other programmes and policies (termination, major redesign, minor redesign, expansion/prolongation of the measure, redesign of another measure, and merger of measures). The consequences reported thus were only those that the policy makers related to the recommendations of the evaluation. In general, 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 redesign of measures or their prolongation and

extension. In 17% of all cases they are also used to improve other or future policy measures. We come back to take a closer look at the determinants of consequences below (Section 5.2).

4. Form follows function? Determinants of evaluation

The next step takes a closer look at the connections between the different dimensions of evaluations. The aim is to better understand if and how the evaluators and commissioners of evaluation design and implement evaluations in order to best fit the defined functional need. This analysis focuses on three key pillars for a functional approach: the link of topics and methods, the meaning of the policy measure evaluated, and the purpose of the evaluation as being mainly formative or mainly summative.

4.1 The link of topics and methods

A first pillar of functional fit is the link of topics and basic evaluative approaches/methods applied: do evaluators tailor their approaches according to the need for topics and impacts to be covered? According to the statistical analysis, this question can be answered with a 'yes, to some extent'. The analysis confirms that there are different groups of topics that are more likely to be analysed with a specific combination of methods and data collection approaches.

A first set of evaluations is concerned with policy and strategy development issues. These evaluations look at external and internal consistency as well as coherence/complementarity. They use context, document, and network analysis significantly more often as well as before/after group comparison. Consequently, they are moderately 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 evaluations of the overall effectiveness of policies (output, outcome, impact) rely on a mix of existing material and rather simple methods to be applied as a standard approach. Evaluations that tackle the overall goals of measures tend to employ case studies, input/output analysis, and descriptive statistics. The data collection methods correlating with this cluster of topics are existing databases, monitoring data, interviews, and participant surveys. In addition, in particular for the assessment of the quality of short- and long-term results and for holistic judgements, peer review, and expert knowledge appear to be of key importance.

A more complex concept for the measurement of impact is additionality, which is differentiated into input additionality (more resources are allocated to innovation activities than would have been without the measure), output additionality (more innovation output), and behavioural additionality (persistent change of actor behaviour that is conducive to better innovation performance) (Gök 2010; Gök and Edler 2011).⁹ In general, evaluations that apply these three types of additionality apply very similar methods and data collection approaches. Compared with evaluations that do not consider additionality aspects, they more often apply econometric analysis, network analysis, and a counter-factual approach. Moreover, evaluations looking at input and output additionality also employ input/output analysis, before/after group comparison, control group, and cost/benefit approach significantly more frequently. The data collection methods used for additionality topics are mainly surveys (either with non-participants or participants and also pre-existing surveys), monitoring data, interviews, and document search. In sum, evaluations concerned with additionality employ analysis and data collection methods that are considered appropriate to the very concepts of additionality.

Evaluations that focus more on programme and project efficiency issues clearly follow a qualitative approach. For those, case studies and context analysis are important, linked with document search, focus groups, and workshops—as it is essential to understand management structures, processes, and practices. Quite logically, efficiency at the project level 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 and peer-review for their technological content.

4.2 The meaning of the policy measure

A second pillar for a functionalist analysis of evaluation assumes that there is some systematic differentiation between the nature of an evaluation and the nature of the policy measure it evaluates. However, there is a strong degree of convergence of evaluation practice across different policy measures. We find surprisingly little variation between different policy measures as regards a whole range of evaluation characteristics, such as tender procedures, internal versus external evaluators, coverage of topics and impacts, use of some of the data collection approaches and methods, and targeted audiences.

However, evaluations often apply tailored methods and data collection approaches (e.g. network analysis and case study approaches for networking and cluster programmes) to meet the specific requirements of complex programmes. We also find some variation in the use and dissemination of evaluation results between policy measures (e.g. in complex networking programmes beneficiaries are much more often informed about the evaluation as these

measures are complex and need explanation and feedback. Furthermore, evaluations of direct financial support measures and of cluster, technology transfer, and networking measures are more likely perceived as being of good quality, while evaluations of softer measures such as management support measures or diffusion measures are perceived to be of lower quality. This, it appears, is due to a much less tangible nature of evaluations and impact attribution for diffusion measures and management measures. In addition, there seems to be a poorly developed evaluation practice for measures that seek to support the diffusion of innovation (see also Arnold and Guy 1997; Edler et al. 2009). Evaluation of those measures do also not take societal and environmental impacts into account as broadly as might be expected, and they are perceived to be less useful to policy makers.

4.3 Learning and assessing: the meaning of the core purpose

Digging a bit deeper, we can focus on the meaning of the specific purpose of evaluations as being largely formative or summative (see also Chen 1996; Patton 1996). As we have shown above, the two purposes do not seem to be entirely distinct. We will come back to this in more detail. In order to sharpen the distinction between the natures of evaluations that are predominantly used for formative purposes against those that help to make a judgement, the first question is: are formative and summative evaluations very different in terms of the topics they cover and the methods they employ?

Two decades ago, Scriven noted that formative evaluations are ‘typically conducted during the development or improvement of a program or product (or person, and so on) and for in-house staff of the program with the intent to improve. The reports normally remain in-house; but serious formative evaluation may be done by an internal or an external evaluator or preferably, a combination; of course, many program staff are, in an informal sense, constantly doing formative evaluation’ (Scriven 1991). Scriven’s distinction has shaped the debate in the evaluation literature and led to many other classifications of the purposes of evaluations. For instance, Mark et al. (2000) argued that what Scriven (1991) labels as summative is ‘assessing merit and worth’ and formative corresponds to ‘program and organizational improvement’. They also suggest that ‘oversight and compliance’ and ‘knowledge development’ are among the possible purposes of evaluation as well. Chen has also questioned this simplistic notion of formative versus summative and has developed a more sophisticated typology which links the formative–summative dichotomy to a second one which is outcome—and process oriented (Chen 1996, see Table 1).

This typology is a conceptual one, built on two dimensions. Its major value for our discussion lies in highlighting that there is no simplistic dichotomy between summative

Table 1. Conceptual evaluation typology following Chen (1996)

	Formative ‘improvement’	Summative ‘judgement’
Process	Process improvement evaluation	Process assessment evaluation
Outcome	Outcome improvement evaluation	Outcome assessment evaluation

evaluations being mainly concerned with outcome, but also with assessing processes. Equally, formative evaluations not only try to improve the process, but by doing so seek to improve outcome as well. The difference between formative and summative thus is not process versus outcome, but judgement versus improvement.

This would then mean that formative evaluations are more about understanding processes and outcomes, i.e. they would be more concerned with consistency and complementarity, project, and programme implementation issues and affect dimensions such as behavioural additionality (learning of changes in behaviour), and that in order to do so, they employ more qualitative methods.

Our statistical analysis shows that formative evaluations indeed cover significantly more often topics such as ‘policy/strategy development’, ‘internal’ and ‘external consistency’ as well as ‘programme implementation efficiency’ and significantly less often input and output additionality. This highlights that the very function of formative evaluation is about understanding the overall fit of a programme to its policy context, and the internal logic and efficiency of the programme it is supporting (see Table 1). Consequently, formative evaluations are less about concrete, tangible additionality. This is consistent with the methodological approaches that are applied. Formative evaluations use ‘input output analysis’ and ‘counter factual’ and ‘control group approach’ significantly less often. Formative evaluations seem to have no ‘unique’ combination of methods but rely slightly more often on document analysis and descriptive statistics and, in general, they tend to lean towards qualitative (document analysis) and interactive methods. Formative evaluations, which are done to support the implementation of the policy measure, are significantly more often done in evaluations that are commissioned by an external sponsor than summative evaluations. This means that evaluations that are done to support the implementation process of the policy measure are more often commissioned from units and agencies outside the implementation unit itself.

Evaluations which are (at least partly) summative are 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 that the results of summative evaluations, with

clear ‘numbers’ and simple messages, are better suited for wider discussion and the demonstration of legitimation. The virtue of formative evaluations is not so much their dissemination, but the fact that they support learning within the process itself, they are a tool for improvement for all parties involved in the programme.

5. Quality and consequences of evaluations

5.1 Determinants of quality

The nature of quality of evaluations above was discussed in Section 3.2. Overall, 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.

A deeper statistical analysis of what influences the policy maker’s quality assessments reveals an interesting differentiation: evaluations covering technological and scientific impact are perceived to be of higher quality than those which examine societal and environmental impact. The latter are obviously much harder to determine, to operationalize, and to attribute to a specific policy measure, and thus findings on those impacts are met with higher scepticism. In terms of methods, evaluations using survey methods and peer review are perceived to be of higher quality. This confirms that there is a general, often unquestioned, belief in survey data and in expert judgement.

Interestingly, perceived 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 pre-designed elements of policy measures and have a dedicated budget. Related to that, evaluations that are done for policy measures sponsored by external (national or international) (co-)sponsors are perceived to be of lower quality. The interpretation is not straightforward; evaluations may be 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 imposed by the external sponsor and do not fit the needs of the specific context. Whatever the reason, there appears to be room for improvement in the design and conduct of evaluations of co-sponsored measures. This is clearly confirmed by the in-depth study on portfolio and structural fund evaluations (Amanatidou and Garefi 2011).

Furthermore, quality perception is related to the tender process. Open tenders yield evaluations with better perceived quality compared with closed or restricted tenders. This is highly significant, it shows that broad competition and the search for the best expertise on the market leads to better evaluations, the excellence provided by the market is more important than context knowledge of those closer to the commissioning body (closed, restricted tenders).

Finally, perceived quality makes a difference when it comes to the dissemination and exploitation of evaluations. Higher perceived quality of an evaluation is correlated with more discussion within and outside government. In addition, evaluations that are targeted to the wider public and policy analysts (and not only to the programme management) are also correlated with higher quality.

5.2 Determinants of consequences

As stated above, an important finding is that the recommendations of evaluations rarely lead to more radical consequences (such as the termination of programmes), as radical shifts appear to be the consequence of more general, principle policy decisions.

An important question in regard to the effects of evaluation is if their consequences are dependent on the perception of their quality. Table 2 displays the correlation between individual quality indicators and consequence categories.¹⁰ A first, important observation is that very

Table 2. Correlation coefficients between individual quality assessment indicators and consequences of the evaluation (Spearman’s, pairwise)

Consequence quality aspect	Termination	Major redesign	Minor redesign	Expansion/ prolongation	Redesign of another measure	Merger of measures	Number of cases
Address TOR	-0.0847	-0.1114	0.0902	0.1648	0.1178	-0.2122*	86
Design appropriate given the objectives	-0.0486	0.0132	0.2134*	0.2382*	0.0805	-0.0505	124
Methods satisfy the TOR/purpose	-0.1441	-0.0653	0.2081*	0.1555	-0.0754	-0.1005	100
Application of qualitative methods	-0.135	-0.0058	0.1325	0.2999*	0.0913	0.0145	117
Application of the quantitative methods	-0.1413	-0.1457	0.1047	0.2952*	0.0733	-0.1129	109
Information sources well documented	-0.1125	-0.1621*	0.1446	0.058	0.0841	-0.0562	125
Analysis based on given data	-0.2098*	-0.0534	0.1627*	0.2884*	0.1396	-0.1141	125
Cover broader context	-0.1589*	-0.0531	0.1478	0.2800*	0.0533	-0.1123	123
Conclusions based on analysis	0.0473	0.0824	0.1794*	0.2403*	0.0377	-0.0245	86

Bold type indicates statistically significant difference at *10% level based on Spearman’s correlation coefficient.

strong changes, such as the termination of a programme or a major redesign, are not linked to the actual quality of evaluations. Major decisions on redesign or termination may depend less on evaluation results and quality than on other considerations such as a change of policy priority. However, it is important to point out again that there are very few cases in which an evaluation caused a termination. Second, evaluations with a higher (perceived) quality have a significant positive 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 recommendations (which might result in extension/prolongation) is more often assessed by policy makers to be of good quality.

The most influential quality aspects are satisfactory methods in relation to the objectives of the terms of reference and the initial purpose. The lesson here is crucial: evaluations may be creative, add new questions, and design new approaches, but in the end, they only convince if they manage to satisfy the initial purpose of the evaluation. Only rarely does the evaluation process itself lead to a change of the terms of reference and thus the expectations of policy makers towards the evaluation.¹¹ Finally, the credibility of an evaluation that is needed for subsequent policy implementation is closely linked to the application of methods, both quantitative and qualitative. The scope of the consequences is highly correlated with simple methods, i.e. clarity and simplicity of the data and its collection and analytical methods are essential for turning the recommendations of evaluations into action. Only if the techniques are appropriate and understandable, the evaluation can convince its audience.

In addition, the nature of the consequences differs for summative and formative evaluations. Summative evaluations, which are perceived as being of higher quality and which are more broadly discussed across government, tend to lead to more severe consequences such as termination, major redesign, or merger of measures. Also, summative evaluations more often lead to expansions of programmes, while formative evaluations tend to cause more often minor modifications and prolongation of the measures they evaluate. This suggests that in order to terminate or radically alter a measure, some quantitative, easy to communicate summative evaluation results are required.

Consequences of evaluations are further related to the way in which evaluation results are discussed across government and stakeholders. There are stable positive relations between the intensity and scope of the discussion about a specific evaluation on the one hand, and consequences on the other. This is true for consequences overall and for the two most frequent consequences minor redesign and programme extension and prolongation (Table 3). Only the less frequent consequences

Table 3. Correlation coefficients between discussion indicators and consequences of the evaluation (Spearman's, pairwise)

Consequences of evaluation	Discussed within government (<i>n</i> = 98)	Discussed with participants/ stakeholders (<i>n</i> = 103)
Termination	-0.0673	0.0129
Major redesign	0.1075	-0.0362
Minor redesign	0.2340*	0.2116*
Expansion/prolongation	0.3229*	0.3454*
Redesign of another measure	-0.1478	-0.0293
Merger of measures	0.1055	-0.001
Any consequence	0.3683*	0.2926*

Bold type indicates statistically significant difference at *10% level based on Spearman's correlation coefficient.

(termination, redesign of another measure, and major redesign) are not linked to the mode of discussion. This points to the need for a translation process, in which evaluation results are dealt with a broader policy context.

6. Types of evaluations

In a final step we performed a hierarchical cluster analysis in order to highlight common features, reduce complexity, and explore evaluation types. We worked with the sample of cases for which policy makers had complemented the data in sufficient detail and quality. This reduces the overall number of cases to 84, as we have to exclude all reports that have a missing variable which is of importance for the analysis. For the cluster analysis, all variables are either binary or ordinal variables that have been transformed into (several) binary variables. For the analysis we used the Ward algorithm which is associated with very positive features (cf. Bergs 1981; Hands and Everitt 1987). The simple matching binary similarity coefficient¹² was used as a distance measure (Finch 2005; STATA 2007). On basis of the Duda/Hart Je(2)/Je(1) (which should be high) and pseudo-*T*-squared values (which should be low and lower than those values of neighbouring cluster numbers) as well as Calinski/Harabasz pseudo-*F* values (which should be high) (see STATA 2007: 154), we established three clusters of which the first includes 53, the second 11, and the third 20 cases. Despite the existence of these quality criteria, it is important to remember that cluster analysis is a highly explorative analytical tool with many possible outcomes which aims at finding groups in data and which is rather intended for generating than for testing hypotheses (see Kaufman and Rousseeuw 1990; Everitt et al. 2001).

The cluster analysis groups evaluations based on information about their characteristics and characteristics of

Table 4. Qualitative summary of cluster profiling

Evaluation dimensions	Cluster 1: the support	Cluster 2: the verdict	Cluster 3: the holistic
Timing	Interim (68%)	Ex post (82%)	Ex post (75%)
Purpose	Formative (68%)	Summative (64%)	Both (70%)
Planning	Foreseen and planned (85%)	Less often foreseen and planned (46%)	Foreseen and planned (85%)
Conducted by	External (98%)	External (46%), but also internal (36%) or mixed (18%)	External (90%)
Topics	Programme efficiency (85%) (and thus management) focused, also consistency (83%/87%) coherence/complementarity (74%) and policy/strategy development (74%) important	Target few topics: mostly output (64%) and goal attainment (55%), also some input additionality (55%); not about internal (9%) or external (0%) consistency, project implementation efficiency (9%) or policy/strategy development (9%)	Target many topics: esp. goal attainment (100%), output (100%) and quality (80%); all types additionality (90%), but also consistency (external (80%)/internal (70%)), programme implementation and policy/strategy development (70% each)
Impact	Impact assessment important (89%), but only considers technological and economic impact (about 50% each)	Impact assessment less often used (64%) but still most important topic. Mainly economic impact (half of evaluations)	Clearly focused on impact (100%), all cover economic impact, 75% technological impact, 60% social impact
Methodological approaches and data sources	Qualitative methods and sources important; either interviews (94%) and focus groups (60%) or document (60%) and context (72%) analysis; participant surveys (77%) and existing surveys/databases (68%) used, but descriptive (79%)	Narrow approach, only few methods/sources used; mostly quantitative (econometric analysis (55%), control group (55%), counter-factual (64%)), based on existing data (46%) and participant surveys (46%)	Broader scope: many methods partially used, esp. interviews (90%), participant surveys (90%) and existing data (70%) are all important; but analysis is restricted to descriptive statistics (100%); In addition, case studies (40%), context analysis (55%), input/output (20%), group comparison (before/after; 10%), cost/benefit analysis (35%)
Measure type	Relative higher share of innovation diffusion (40%) and uptake measures (25%)	2/3 are about direct financial support (accountability)	Focus on science-industry cooperation (45%) network (35%) and spin-off (20%) programmes

Note: highlighted are those aspects of a given cluster that are either valid for a large number of evaluations within this cluster (e.g. 85% of evaluations of cluster 1 are foreseen and planned) or aspects which stand out in comparison to one or two other clusters (e.g. only 20% of the evaluations in cluster 3 are focused on spin-off measures, however compared with 4% in cluster 1 and 0% in cluster 2 this is a relatively high share).

the policy measures they evaluate. In detail the following aspects are included:

- evaluation characteristics: timing and purpose, conductor (internal, external), topics and impacts covered, analysis and data collection method applied.
- appraised measure characteristics: modality and target group of measure.

Although by nature of the clustering exercise, there is a lot of overlap and the typification is somewhat stylized, three distinct types emerge, as summarized in Table 4.

The first type is the *supporting* evaluation. This is largely formative, it is planned and thus part of the measure cycle. As a management tool it focuses very much on the consistency and coherence of the programme and its efficiency. It is more hands-on with its methods, largely relying on workshops and group interaction and document searches.

This approach is used for policies supporting diffusion and uptake of innovation as well as for those measures that try to create new supporting structures and intermediate bodies.

An opposite type is largely summative, and it can be labelled 'the verdict' (cluster 2). Those summative-only evaluations are more likely to be ex post, it appears that they are not planned for in the sense of an integrated policy cycle approach, but done for reasons of justification and accountability. Those evaluations are not about process learning, they do not check for consistency and complementarity, policy development or efficiency. In short, they do not take into account the very context of the programmes that are evaluated and the relative focus is on economic impact and input additionality. Following from the quantitative focus, there is an obvious neglect of qualitative methods and data sources. Further, in

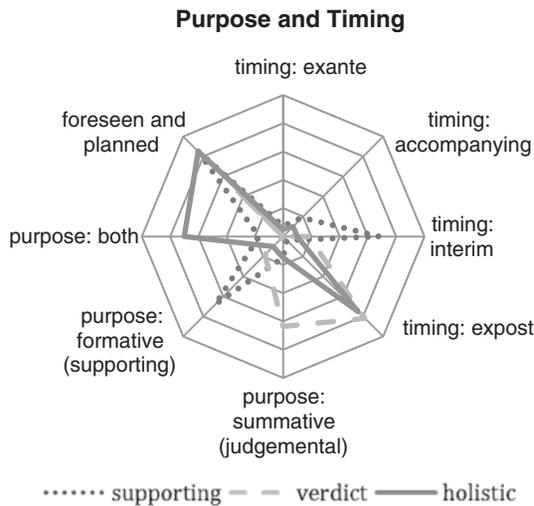


Figure 5. Purpose and timing. Share in percent of type.

comparison to the other evaluation types, the verdict approach is more often conducted for those measures that provide direct financial support.

A third type is a much more *holistic* approach, combining formative and summative elements. Like the purely formative approach it is planned, but it is much broader in its coverage of topics and application of methods. It combines the efficiency approach (on project level even) with the measurement of goal attainment, effectiveness, and a range of impact dimensions. Being holistic also means to understand and measure the programme logic, thus the approach focuses on the integration of all types of additionality in its analysis. As for methods, it focuses on combining survey data with case studies, some (limited) network analysis and applies, as a consequence of the additionality assessment, some (limited) before/after group comparisons. Interestingly, this approach—as with the formative approach more generally, also *relies on peer reviews* (this is even more true for cluster 1: 19% cluster 1 and 15% in cluster 3), and by doing so, it brings in technological and economic expertise in addition to participant surveys and interviews of target groups and management. The holistic approach is especially important for measures supporting collaboration, as here the effects are on various levels and build on each other (e.g. cooperation improvement leading to impacts) and on heterogeneous target groups, and the challenges for management and more generally for assessment are most complex.

We can visualize the characteristics of three types. Figure 5 shows the key characteristics of the three clusters or types of evaluations, while Figure 6 depicts the share of evaluations that cover the various topics and impacts and Figure 7 shows what kinds of policy measures are covered by the types of evaluations.

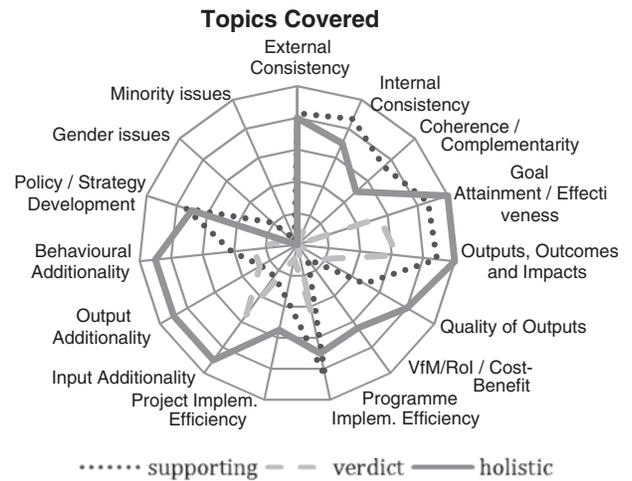


Figure 6. Topics covered. Share in percent per type that cover specific topic.

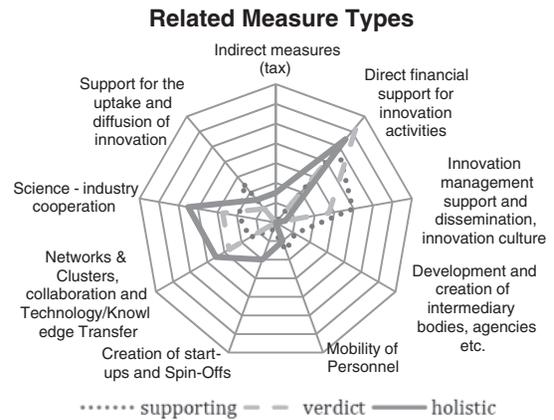


Figure 7. Related measure types. Share (%) of evaluations per type that evaluate specific policy measures.

How do those three stylized types relate to quality, usefulness, and consequences? Policy makers assessed specific quality aspects on the basis of which a quality index was constructed. On that basis, the quality of the holistic approaches is much higher in all dimensions (Figure 8). The verdict evaluation is rated much worse than other types when it comes to the application of qualitative methods, the coverage of the broader context and the fulfilment of the Terms of Reference. This hints towards a mechanistic application of largely quantitative methods. The supporting, purely formative evaluation, is close to the holistic evaluation, but lacks the application of quantitative methods.

Evaluations need to be perceived as useful by policy makers. As explained above, usefulness was defined for a set of pre-selected dimensions. For all evaluations policy makers first indicated if an aspect of usefulness was covered in a recommendation and if so, they indicated

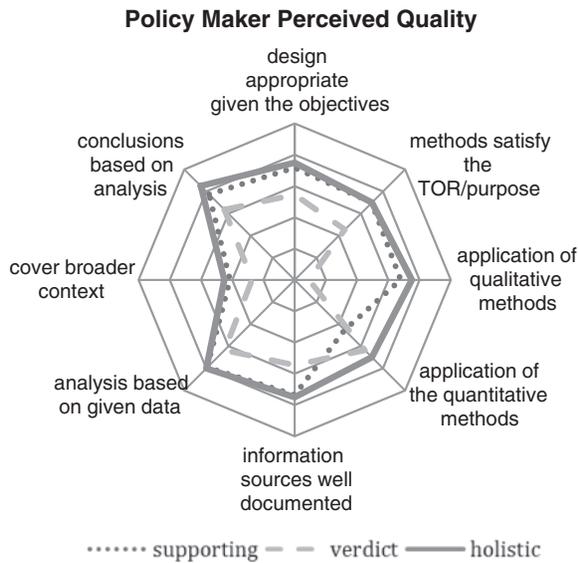


Figure 8. Policy maker perceived quality. Share (%) of evaluations which policy makers rated above 3 on a quality scale 1 = very low to 5 = very high.

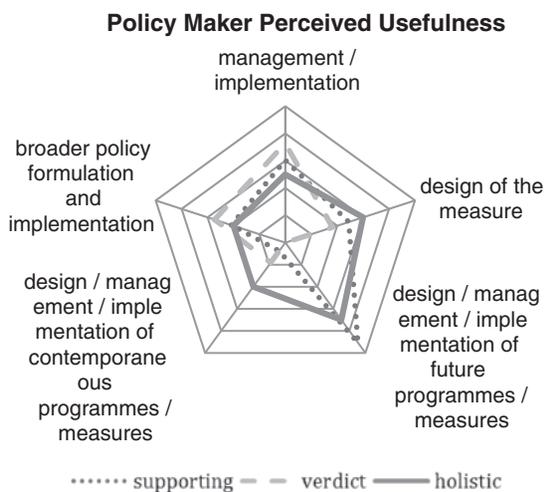


Figure 9. Policy maker perceived usefulness. Share (%) of type which policy makers rated useful above 3 on Likert scale 1 = not useful at all to 5 = very useful.

how useful the recommendation for this aspect was. The Likert scale ran from 1 = not at all to 5 = definitely. Figure 9 indicates that the recommendations based on the holistic approach were far more useful for modification of the design of the evaluated measure as well as design, management, and implementation of other and future programmes. Interestingly, the purely summative was perceived to be more useful for the management and implementation of the evaluated measure and for broader policy formulation. This is in line with the finding reported above that summative evaluations are easier to communicate to government circles, they are essential

when it comes to justifying and deciding about policies, while the holistic learning evaluation spills over to policy design more generally.

7. Conclusions

This article set out to shed light on evaluation practice in innovation policy in Europe based upon empirical evidence. It delivered a conceptualization of evaluations and analysed the phenomenology of evaluations on the basis of the largest existing evaluation report database. Such an analysis is timely, as the requests for high quality, high impact evaluation in a functional fit approach are mounting. In contrast to previous work on evaluation practice in STI (Science, Technology and Innovation policy) policy more broadly, this article focused on the overall picture of evaluation practice, rather than on country comparison, in-depth analysis of specific methods or types of evaluation or on best practice cases. This broad state of the analysis, we believe, is a sound basis from which further, more specific and qualitative analysis can start.

The analysis and the discussions with policy makers in our study indicate that evaluation is more and more an integral part of innovation policy design. (i.e. two-thirds of evaluations have been planned and foreseen, almost half of them had a dedicated budget). There is, however, quite a way to go. Evaluation designs do not yet show the functional fit one would expect. They are not yet fully tailored to the specific characteristics of the programmes they evaluate. While we see some tailored methods being used, the overwhelming impression is that there is a core of methods and approaches applied across a range of measures regardless of the specific needs to cover certain topics or policy programme objectives. This may very well be the major source for the fact that policy makers assessed more than one-third of evaluations as being poor.

One of the key findings relates to the purpose of evaluation as being formative and/or summative and the linkages regarding characteristics of evaluations. First, we can empirically confirm Chen's (1996) view that the expectations of evaluation are often both formative and summative and are both about process *and* outcome. Secondly, there are differences between those approaches that are *mainly* formative or *mainly* summative in terms of methods applied, topics covered, and of who performs and sponsors the evaluation. Formative evaluations are indeed more about context, consistency, and process. Interestingly, however, it is not the formative evaluation that leads to broader discussion and more radical adjustment of policies, but the summative evaluation. This has to do with the 'verdict' element of summative evaluation, often based on a set of simple numbers that can be communicated much easier. At the same time, formative evaluations are more often commissioned by actors outside

the implementing unit, which suggests that often learning is imposed. This means we need to enlarge our perception of what 'learning' means: while formative evaluations contribute to a process improvement and are often sponsored by third actors to do so, the verdict and message of summative evaluations and their interpretations are instrumental for higher level adjustments of policy. Thus, a simple dichotomy between formative and summative, as is often done in evaluation analysis and practice, is not sufficient. It also shows that the evaluations with the highest perception of quality are clearly those that are holistic, combining both summative and formative approaches and purposes.

Further, evaluation reports are more likely to make a difference if they are widely discussed within government and with stakeholders. Thus, the 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. A further conclusion, based on the interviews with policy makers rather than the statistical analysis, is that evaluators themselves have to bear in mind the consequences their evaluations have in terms of influencing the debate on the instrument. Further, the effects of their evaluations also depend on the perceived quality of their reports and the clarity with which methodologies are described and results presented. At the same time, evaluation practice needs close interaction at all stages between those commissioning and those performing the evaluations. To that end, policy makers need to be 'intelligent customers'; they need to have the absorptive capacity to understand what evaluations can deliver and what they cannot deliver. If those conditions are given, the process and the results allow both policy makers and evaluators to reflect on their own practice, their approach to evaluation, and ultimately, the use of evaluation.

Looking ahead, it appears that in future 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 evaluation of innovation policy, and the methods used, will have to adapt to new trends in innovation policy, as policy measures are likely to differentiate even more in the future. Evaluation practice in Europe will have to follow the principle of 'form follows function' much more closely, making the identified need to overcome the uniformity of methods identified in our study even more pressing. Allowing for more experimentation will become more important.

To highlight one key example, one major trend is the increasing policy focus on 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 need to improve radically before adequate evaluation approaches can be developed, and this understanding has to be shared by policy makers and evaluators alike (Edler et al. 2012).

Yet, a dilemma confronting evaluation has to be noted. As just noted, in order to provide new methods and concepts to better inform policy, evaluation itself has to be innovative and experimental. However, 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.

In order to improve the relevance and usefulness of the future analysis of evaluation practice itself, our study has paved some new ground. First, we believe that the very conceptualization of the nature—the phenomenology of evaluations in innovation policy—is an achievement in itself. Although the conceptualization for usefulness and quality used in this study (assessments of policy makers and proxies based upon these) should be further enhanced and complemented by alternative measures, we believe that it is important to know how evaluations are actually perceived by those using them. In any case, this conceptualization can and should be used for further comparative and conceptual work, and it has already served this purpose (e.g. Amanatidou and Garefi 2011; Bühner and Daimer 2011; Gök and Edler 2011). Secondly, the underlying study has established a repository of evaluation data that can be used as reference point for evaluation and policy practitioners as well as for all kinds of future analysis.

The need for further analysis is obvious; the article has raised or reinforced many questions to be further explored. To highlight the most important one: contextualization of evaluation design and usefulness is called for. This is about trying to better understand the context conditions and requirements to make evaluations useful for policy makers. We know that decisions to change, abandon, or design programmes are the result of complex processes and interests and they are often politically driven. Evaluation is one source of legitimization and helps to define problems. Evaluations provide stakeholders with information, analysis, recommendation, and enable reflexivity. On an even more basic level, they act as a focus for discourse on future policies and as a bridging mechanism between programme owners and managers, higher level political decision makers, the beneficiaries, and the wider public. They enable and support policy-oriented discourse, not more, not less. But our knowledge about what enhances the likelihood that evaluations can perform these enabling functions is scarce and based on anecdotal evidence by reflective policy makers (Pichler 2010). Our article has

started to look at connections between perceived quality and usefulness and the consequences of different types of evaluations. This kind of analysis must be linked to contextualized case studies with a focus on the connection between the main aim of evaluations, the design of evaluations, and the political context. This would bridge the gap between the de-contextualized, albeit verified, statistics presented in our article and the anecdotal and few robust experiences we know about.

Acknowledgement

The study that enables this article was sponsored by the European Commission, DG Enterprise. We would like to thank the sponsor for the financial and intellectual support. Details of the study can be found at the INNO-Appraisal homepage <http://www.proinno-europe.eu/page/inno-appraisal>. We also would like to thank the members of the INNO-Appraisal project team, especially Dr Paul Cunningham, for their valuable comments and insights.

Notes

1. In fact, INNO-Policy Trendchart database of policy measures has reached to 1000 policy measures in 2008 from less than 200 in 1995 (Tsipouiri et al. 2008).
2. Edler et al. (2008) give some account of different approaches to meta-evaluation and meta-analysis.
3. The full database used in this study can be accessed at <http://proinno.intrasoft.be/index.cfm?fuseaction=page.display&topicID=262&parentID=52>. Since the conclusion of the data collection, INNO-Policy Trendchart database has been merged with the ERAWATCH database on science policy which can be accessed at <http://erawatch.jrc.ec.europa.eu>.
4. This database has been compiled in the INNO-Appraisal project financed by the EU Commission, conducted by a team from MIOIR, Manchester; Joanneum Research, Austria; Wise Guys, UK; Fraunhofer ISI, Germany; and Atlantis, Greece. It was led by one of the authors, Jakob Edler, <http://www.proinno-europe.eu/page/inno-appraisal> (Edler et al. 2010). The authors of this article thank all partners of this study for valuable support and comments on earlier versions of this article.
5. For the time the study covers Romania and Bulgaria where not yet members of the EU.
6. Twenty percent of the evaluations included in the INNO-Appraisal database is from Austria, 11% Germany, 8% Czech Republic, 7% UK, 6% Netherlands, while the remaining 48% are shared by 20 other EU Member States. Readers interested in the distribution of reports per country can find a graph in Edler et al. (2010: 36).
7. For more information on the particulars of the data collection procedure, see the data collection manual at http://www.proinno-europe.eu/extranet/admin/uploaded_documents/INNO-Appraisal_Data_Collection_Storage_Manual.pdf.
8. The study has confined itself to those four major impact dimensions, based on the assumption that these are the most relevant for innovation policies. We acknowledge that there is a much richer variety of impact (OECD 2009: 149).
9. We follow a widely shared definition of behavioural additionality 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 (Buisseret et al. 1995).
10. It must be noted that for some consequence categories the overall number of cases is rather low.
11. Examples for that can be found in Gök and Edler (2011) in the context of behavioural evaluation studies.
12. Similarity measures for binary data are based on the four values from the cross-tabulation of observations i and j (when comparing observations). Given that a is the number of variables where observations i and j both had ones, and d is the number of variables where observations i and j both had zeros. While the number of variables where observation i is one and observation j is zero is b , and the number of variables where observation i is zero and observation j is one is c . In this case the simple matching binary similarity coefficient is: $(a + d)/(a + b + c + d)$ (STATA 2007: 496).

References

- Amanatidou, E. and Garefi, I. (2011) *Evaluation in the Context of Structural Funds: Impact on Evaluation Culture and Institutional Build up*, EUNIP International Workshop on Evaluating Innovation Policy: Methods and Applications, 5–6 May 2011, Florence.
- Arnold, E. (2004) 'Evaluation Research and Innovation Policy: A Systems World Needs Systems Evaluations', *Research Evaluation*, 13/1: 3–17.
- Arnold, E. and Guy, K. (1997) 'Technology Diffusion Programmes and the Challenge for Evaluation', In: OECD (ed.) *Policy Evaluation in Innovation and Technology*. Paris: OECD Proceedings.
- Bergs, S. (1981) *Optimalität bei Cluster-Analysen: Experimente zur Bewertung numerischer Klassifikationsverfahren*. PhD Thesis. Muenster.
- 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.
- Bührer, S. and Daimer, S. (2011) *The Role of Impact Assessment in Evaluation*, EUNIP International Workshop on Evaluating Innovation Policy: Methods and Applications, 5–6 May 2011, Florence.

- Chen, H. T. (1996) 'A Comprehensive Typology for Program Evaluation', *Evaluation Practice*, 17/2: 121–130.
- Corder, G. W. and Foreman, D. I. (2009) *Nonparametric Statistics for Non-Statisticians: A Step-by-Step Approach*. Hoboken: Wiley.
- Edler, J., Ebersberger, B. and Lo, V. (2008) 'Improving Policy Understanding by means of Secondary Evaluation', *R&D Evaluation*, 17/3: 175–186.
- Edler, J., Blind, K., Georghiou, L. and Uyarra, E. (2009) *Monitoring and Evaluation Methodology for the EU Lead Market Initiative: A Concept Development*, Report for the Innogrips/Lead Market Initiative, DG Enterprise, European Commission, Brussels.
- Edler, J. et al. (2010) *INNO-Appraisal. Understanding Evaluation of Innovation Policy in Europe*, Brussels/Manchester.
- Edler, J., Georghiou, L., Blind, K. and Elvira, U. (2012) 'Evaluating the Demand Side. New Challenges for Evaluation', *Research Evaluation*, 21: 33–47.
- Everitt, B. S., Landau, S. and Leese, M. (2001) *Cluster Analysis*, 4th edn. London: Arnold.
- Fahrenkrog, G. et al. (2002) 'RTD Evaluation Toolbox – Assessing the Socio-economic Impact of RTD Policies (EUR 20382 EN)'. Seville: IPTS. <www.jrc.es/home/publications/publication.cfm?pub=1045>.
- Finch, H. (2005) 'Comparison of Distance Measures in Cluster Analysis with Dichotomous Data', *Journal of Data Science*, 3/1: 85–100.
- Flanagan, K., Uyarra, E. and Laranja, M. (2011) 'Reconceptualising the 'Policy Mix' for Innovation', *Research Policy*, 40: 702–713.
- Georghiou, L. (1999a) 'Meta Evaluation. Evaluation of Evaluations', *Scientometrics*, 4/3: 523–530.
- . (1999b) 'Issues in the Evaluation of Innovation and Technology Policy', In: OECD (ed.) *Policy Evaluation in Innovation and Technology Policy: Towards Best Practice*, pp. 19–33. Paris: OECD.
- Gök, A. (2010) *An Evolutionary Approach to Innovation Policy Evaluation: Behavioural Additivity and Organisational Routines*, PhD Thesis. Manchester Institute of Innovation Research, University of Manchester, Manchester. <<https://www.escholar.manchester.ac.uk/jrnl/item/?pid=uk-ac-man-scw:98600>>.
- Gök, A. and Edler, J. (2011) *The Use of Behavioural Additivity in Innovation Policy-Making*, EUNIP International Workshop on Evaluating Innovation Policy: Methods and Applications, 5–6 May 2011, Florence.
- Hands, S. and Everitt, B. (1987) 'A Monte Carlo Study of the Recovery of Cluster Structure in Binary Data by Hierarchical Clustering Techniques', *Multivariate Behavioral Research*, 22/2: 235–243.
- 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, Brussels.
- Kaufman, L. and Rousseeuw, P. J. (1990) *Finding Groups in Data: An Introduction to Cluster Analysis*. New York: Wiley.
- Mark, M. M., Hentry, G. T. and Julnes, G. (2000) *Evaluation: An Integrated Framework for Understanding Guiding, and Improving Policies and Programs*. San Francisco: Jossey-Bass.
- Miles, I. et al. (2005) *SMART Innovation: A Practical Guide to Evaluating Innovation Programmes*, A Study for DG Enterprise and Industry, Brussels.
- OECD. (1998) *Best Practice Guidelines for Evaluation*, PUMA Policy Brief No. 5, May 1998 <<http://www.oecd.org/dataoecd/11/56/1902965.pdf>>.
- OECD (ed.) (1999) *Policy Evaluation in Innovation and Technology Policy: Towards Best Practice*, Paris.
- OECD. (2009) *Enhancing Public Research Performance Through Evaluation, Impact Assessment and Priority Setting*, OECD DSTI/STPTIP(2009)5, Paris.
- Papaconstantinou, G. and Polt, W. (1999) 'Policy Evaluation in Innovation and Technology Policy. An Overview', In: OECD (ed.) *Policy Evaluation in Innovation and Technology Policy: Towards Best Practice*, pp. 9–14. Paris: OECD.
- Patton, M. Q. (1996) 'A World Larger than Formative and Summative', *Evaluation Practice*, 17/2: 131–144.
- Pichler, R. (2010) *Usefulness of Evaluations*, Presentations at the CIA4OPM OMC-Net Meeting, Brussels, December 2010.
- Řezanková, H. and Everitt, B. (2009) 'Cluster analysis and categorical data', *Statistika*, 216–232.
- Ruegg, R. and Feller, I. (2003) *A Toolkit for Evaluating Public R&D Investment, Models, Methods, and Findings, from ATP's First Decade*. Gaithersburg: NST.
- Scriven, M. (1991) 'Beyond Formative and Summative Evaluation'. In: McLaughlin, G. W. and Phillips, D. C. (eds) *Evaluation and Education: A Quarter Century*, pp. 19–64. Chicago, IL: University of Chicago Press.
- Shapira, P. and Kuhlmann, S., eds, (2003) *Learning from Science and Technology Policy Evaluation*. Northampton, MA and Cheltenham, UK: Edward Elgar Publishers.
- STATA. (2007) *Multivariate Statistics*. College Station: Stata Press.
- Tsipouri, L., Reid, A. and Miedzinski, M. (2008) *European Innovation Progress Report 2008*. Brussels: EU DG Enterprise.