Fraunhofer Institute
Systems and
Innovation Research

Fraunhofer ISI Discussion Papers *Innovation System and Policy Analysis,* No 12/2006
ISSN 1612-1430
Karlsruhe, December 2006

Improving Policy Understanding by Means of Secondary Analyses of Policy Evaluation.

A concept development[1]

Bernd Ebersberger, Jakob Edler, Vivien Lo

Karlsruhe, Germany

---

1 This Working paper is the result of a project funded by the Strategy Fund of the Fraunhofer Institute for Systems and Innovation Research.

Content

1 Introduction

The design and re-design of research, development, technology and innovation (RTDI) policy is complex. Appropriate policymaking needs a broad knowledge base about context conditions, group behaviour, instruments and their mix and, last but not least, policy effects. To provide this knowledge, a broad web of distributed strategic intelligence (Kuhlmann et al. 2001) has been established in OECD countries. One cornerstone of this knowledge-providing system are evaluations of policy measures. Evaluations are used to inform policymakers, programme managers and other stakeholders about the effectiveness, efficiency, appropriateness and impact of a policy measure. This is done to assess past performance (summative) and/or to assist policymakers in the design, implementation and re-adjustment of policies (formative). Stemming from the rationale that policymaking needs systematic information, hundreds of evaluations have been conducted in the OEDC world, and each day the number grows.

In this paper we argue that the existing evaluations are much more than helpers to judge and improve individual, specific policy measures. Rather, these existing evaluations can – and should – be used more systematically to learn more, to improve our understanding of policies beyond the individual cases targeted in any given evaluation. We suggest a concept of systematically using and exploiting existing evaluations for the purpose of learning on the individual, programme and systemic level. This concept serves two main purposes:

(1) to permit better comparison and understanding of measures and their effects by taking into account the large number of observations already gained from existing evaluations (by means of a meta-analysis),

(2) to assess – in a systemic understanding of policymaking – the overall combined effects and the remaining bottlenecks and redundancies of policy measures in a systems world (by means of a modified evaluation synthesis).

The first of these two purposes is obvious. In principle, evaluations are individual case studies of policy measures. However, the information given in a large number of such evaluations can be so uniformly processed that the relationship between programme variables and their effects in qualitative and quantitative terms can be analysed. Through combining a larger number of evaluations, the number of observations is multiplied, and individual cases are transformed into a larger data set. This data set can then be used to assess effects of certain types of policy measures, design variables or context variables, on the one hand, and effects on the other hand. Such analyses will

permit much more systematic comparison of policy measures and – more broadly – more general statements to be derived on the level of individual interventions. Such an exercise can never take the contextual situation of each and any evaluation used fully into account. However, the basic idea is to learn and compare through uniform transformation of individual data sets into one larger data set. A meta-analysis thus complements econometric approaches that build upon a large set of innovation data that happens to include information on support activity.. The advantage of the meta-analysis is that the data collection and preparation enables much more specific and deeper analysis of policy instruments. The challenge of a meta-analysis, however, is to collect and prepare the data in a way that allows a variety of evaluations to be transformed into such a data set.

The second purpose is often referred to, but rarely followed up in systematic studies. It relates to the systems perspective in RTDI policymaking and policy analysis that calls for an evaluation approach that can tackle policy mixes, policy interplay and systemic policy effects. The approach we propose is a modified evaluation synthesis. Evaluation synthesis is – in its most basic definition – an aggregated content analysis based on multiple evaluation reports on similar programmes or projects (Beywl & Associates 2004). In its original meaning, evaluation synthesis was applied to better understand one treatment or programme by way of synthesizing a number of studies done on these individual treatments or programmes. Our modification lies in the fact that we apply this method to a set of programmes in the area of RTDI policies with the main purpose not to better understand one single programme, but to shed light onto the interplay of programmes (policy mix) at the level of innovation systems. Furthermore, we focus on the qualitative aspect of such a synthesis, as we need to synthesise a variety of measures and evaluations in a complex web of interrelations situated in specific system contexts.

The remainder of this paper is organised as follows.

Chapter 2 starts with the principle design of the secondary analysis. Here we discuss the principle directions such a secondary analysis can take and the various consequences this implies for the design of the study approach.

Chapter 3 contains the basis for both the meta-analysis and the evaluation synthesis, i.e. a sound stock-taking and characterisation of existing evaluations and policy measures that are of interest for a given analytical task in the in-depth or in the policy mix analysis. The cornerstone of such a collection and qualification is a "meta-evaluation" to assess the character and quality of the evaluations that will be the input for the two variants of secondary analyses. While various concepts of meta-evaluation exist, we

follow Widmer (1996), Stufflebaum (2001, 2002) and Cooksy/Caracelli (2005) in defining meta-evaluation as an "evaluation of evaluations". In our concept, meta-evaluation will be used in a pragmatic, checklist type of approach in order to prepare the two analytical tasks of the meta-analysis and the evaluation synthesis. While using meta-evaluation to prepare evaluation synthesis has been proposed earlier (Cooksy/Caracelli 2005), we propose to broaden the concept, and to use meta-evaluation as preparation both for meta-analysis and evaluation synthesis and, furthermore, to take advantage of this method in order to assess evaluation culture and capacity (in a given system; see Exhibit 1 below).

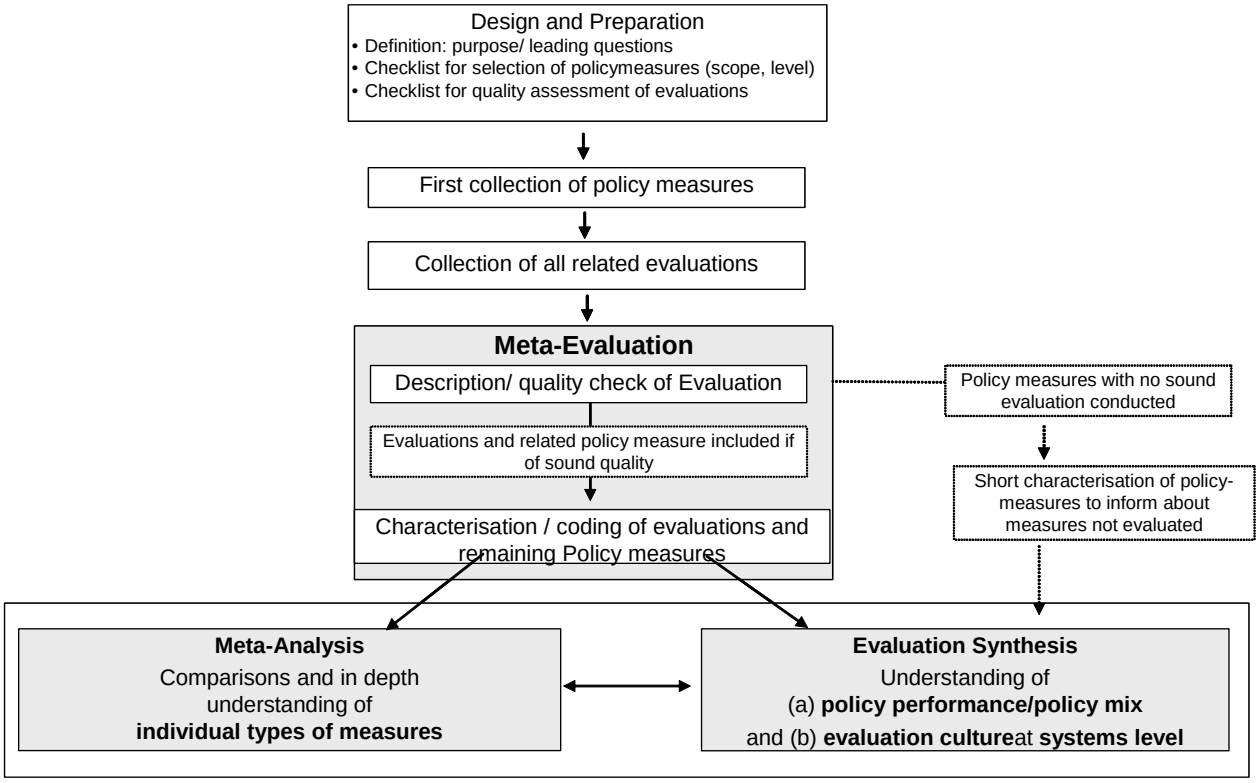
In Chapter 4 we will lay out the principle approach of the in-depth analysis by means of a meta-analysis. The chapter will show the principle ideas of such a meta-analysis and what methodological steps have to be taken in order to use this method for the purpose of policy-learning in the area of RTDI policy.

The major research questions to be tackled with the evaluation synthesis and some methodological principles of this approach are presented in Chapter 5. This chapter will concentrate on the systems perspective, and it will ask what the methodology can contribute to the puzzle of policymaking in complex systems, including a short discussion of the possibilities to assess evaluation culture and capacities.

Finally, chapter 6 summarises the benefits and hurdles of such complex secondary analysis in RTDI policymaking.

Exhibit 1 presents a rough overview of the various elements of this concept.

Exhibit 1: The Concept of Conducting Secondary Evaluation Analysis



Source: Amended and largely modified version of Cooksy/Caracelli (2005, p. 33).

2 Meta-evaluation to Systemise the Input

2.1 Definition

We use meta-evaluation in this concept strictly in the meaning of "evaluating evaluations" (Scriven 1991, Widmer 1996, Stufflebaum 2001, 2002, Cooksy/Caracelli 2005), to assess the quality, relevance, effects, and usage of evaluations (Widmer 1996). The general purpose of these meta-evaluations can be both formative, i.e. assisting policy-makers and evaluators to learn about the evaluation capability and usage and to adopt evaluations according to results obtained and discourses started, and a summative one, i.e. to judge the benefit and quality of evaluation activities. A second meaning of the term in some of the literature, i.e. to combine, aggregate, compare content of evaluations is not within our definition. This clear distinction is needed, as we further differentiate the usage of content in what we have labelled secondary analysis (meta-analysis, evaluation synthesis). In our concept, meta-evaluation serves as a preparation for and core element of the secondary analysis, as the assessment of evaluations is needed in order to judge if the results of these evaluations can be used as input for

the secondary analysis. Meta-evaluation also plays a prominent part when it comes to assessing evaluation culture itself as an element of policymaking in innovation systems (see chapter 4.3.4).

2.2 Starting Points: Clear Target Definition and Preparation of Data Coding

There are two guiding principles: (1) the selection of measures and evaluations has to follow the overall idea of the whole secondary analysis. Neither is it necessary nor fruitful to maximise the inclusion of measures and evaluations, rather they must fit the study purpose. (2) The information given in evaluations shall be transformed into data sheets, if possible and practicable as a mixture of numerical data code and text code in addition.

Selection follows target

The secondary analysis seeks to compare and contrast measures and policy mixes through either qualitative techniques (e.g. inspection of the range of profiles and allocation to empirically determined categories), or via the adoption of more quantitative techniques. In a first step, the evaluations and measures need to be characterised and assessed.

Both policy measures and evaluations have to undergo the systematic screening, selection process and analysis (on the basis of the meta-evaluation). The overriding principle is that the goals of the secondary analysis determine the selection of policies and evaluations.

The major task at the beginning of a secondary analysis in our understanding is to clearly define the overall purpose. For the meta-analysis that deals with individual measures, this means that all policy measures shall be included that contribute to the attainment of the goal that is of interest for the meta-analysis in the first place. If the goal of the meta-analysis is to understand which policy measures best contribute to a given policy goal, what design characteristics of these programmes work best, and how the programme interact with contextual conditions, all measures that tackle these specific goals (explicitly) would be selected.

For the evaluation synthesis, the scope of analysis needs to be clearly defined. What are the system boundaries, how are these boundaries defined, who are the target groups for the measures to be selected and how are these target groups influenced by policies outside the innovation system defined? What are the system functions to be addressed in the analysis, comprehensive or limited to certain functions (and policy

objectives); if so, which ones? All these questions have to be clarified before the selection process is started.

Coding and assessing

The basic idea of preparing secondary analysis is to transform the characterisation of a potentially high number of measures and evaluations into a data sheet via a systematic characterisation and coding scheme. The classification is done for the phenomenology of evaluations and policies as well as their quality (performance). For both levels, a set of criteria will be defined. This should follow a scheme that needs to be defined ex ante, but can be modified along the process of secondary analysis. A starting point for the characterisation of evaluations (3.1) and policy measures (3.2) is given below.

The coding can be done by allocating numbers to a specific criterion. For example, when coding the target function of a policy measure, a list of potential targets for a type of policy measures one can be developed and binary codes allocated to each of the targets. Alternatively, each target can be given a number and the data sheet then includes all the relevant numbers for the target of a specific policy. The number-coding mainly feeds into the meta-analysis, but can also be used for the basic screening within the evaluation synthesis. However, especially for the evaluation synthesis, the coding should be accompanied, wherever needed and feasible, by a text field that gives additional qualitative information and specifies the criteria.

While the positive criteria, i.e. the criteria that simply describe and typify, are rather simply coded, coding is more challenging for the criteria that assess the quality or performance of a measure or evaluation. For the evaluation, all categories within the "quality" dimension are relevant, for the policy measure both the performance categories as well as the categories regarding role of evaluation are of relevance. For each of these categories, a positive benchmark needs to be defined, a standard measure, accompanied by nominal or categorical data as needed. Then, a score based on the results of the evaluations (for policy measures) and of the researcher conducting the secondary analysis (for the evaluations) can be applied. The scorecard approach can be used to aggregate performance at the dimension level to say something about performance across policies and, ultimately, across profiles constructed at systems level (e.g. national and regional levels).

2.3 Characterisation of Evaluations and Evaluation Culture

In our concept, the main purpose of the meta-evaluation is to prepare for and assist the following secondary analyses. It serves to characterise the evaluations and acts as a quality filter for the selection of policy measures. Both functions are intertwined. Only

those policy measures and evaluations can be considered for further analysis that pass the quality check of the meta-evaluation. In assessing the quality of evaluations, a checklist of quality criteria must be followed. This qualification is a major part of the characterisation process of the evaluation study in the first place.

Meta-evaluation is a method of assessing the quality of evaluations (Cooksy / Caracelli 2005). In a broader definition, Stufflebaum has defined meta-evaluation as the process of delineating, obtaining, and applying descriptive information and judgemental information about the utility, propriety, and accuracy of an evaluation (Stufflebaum 2002, p. 95). For our purposes, a set of evaluations will have to be assessed for feeding into the secondary analysis. Thus, the concept of a meta-evaluation cannot be too rigid, comprehensive or ambitious. Rather, checklists of (1) typification and (2) quality shall suffice.

2.3.1 Typification of Evaluations

The checklist for characterising evaluations reflects the variety of evaluation approaches and purposes. It serves to characterise the nature of the evaluation according to a number of dimensions. This is needed to ascertain the value of an evaluation and its results. Furthermore, it enables quantitative and qualitative analysis regarding the connection between policy and policy performance. Finally, the typification increases our understanding of the range of evaluation types, the dominant modes in use and the different combinations of evaluation types, evaluation functions and policy measures deployed in different national and regional contexts.

To typify evaluations, the following dimensions and criteria can be used:

- Time of evaluation: ex ante, accompanying, interim, ex post
- Type of addressee of the evaluation (programme manager, policymaker) and stakeholder involvement
- Aim and purpose of the evaluation: formative, summative, justificatory, routine (as part of programme monitoring and reporting) or ad hoc, continuation review
- System perspective in evaluation: discussion of role of the programme in the system or portfolio, singular/isolated approach
- Method mix: participant/ context interviews, expert panels – peer review, input analysis, cost – benefit, econometric analysis survey-based, econometric analysis based on innovation data (e.g. CIS), case studies, network analysis, end-user analysis, outputs analysis (bibliometrics. patents), etc.
- Composition of the study team: independent/internal/affiliated, academic/private/public sector, self-evaluation of the programme management, etc.

- Evaluation process organisation: transparency, openness
- Recommendations: presence/absence, scope, content (minor adjustment, rigorous changes, ending a programme etc.)
- Evaluation issues addressed: uptake, efficiency, effectiveness, additionality, value-added, impact, gender issues, etc.
- Role and impact of the evaluation in the measure cycle: dedicated evaluation budgets, presence of ex ante evaluation, development of programme models, logic charts etc, in programme design, development of evaluation planning, role of monitoring arrangements, level (purpose) of evaluation conducted (i.e. which factors/processes are assessed), impact and utility of the evaluation for the subsequent policy process, feedback and integration with policy formulation.

Stemming from this characterisation and assessment exercise, however, there is a second potential benefit in conducting meta-evaluation and subsequent evaluation synthesis, i.e. learning about the conduct of evaluations and assessment of the evaluation culture in any given innovation system. The role of evaluation in itself is a quality criterion for a policy measure in the secondary analysis. The quality and rigidity of the evaluations in a given country is an indication of the evaluation culture and as such for the nature of innovation policy governance, signalling a reflexive, evidence-based mode of governance in innovation policy.

2.3.2 Quality of the Evaluation

While the evaluations are described and coded along the various dimensions mentioned in the classification process, a second step is to assess critically the **quality** of the evaluation. Quality is not defined in a global consensus of evaluators. We develop a checklist for quality criteria on the basis of three sources, from which we extract those criteria that are most relevant and useful for our specific purpose of checking evaluation quality for secondary analysis. This checklist draws on existing standards of evaluation and on additional criteria developed in an on-going study for the European Commission (Edler et al. 2006a).

First, there is a common understanding of a set of principles to be followed when conducting evaluations. These principles are first general ones, such as standards of various evaluation societies.[2] The evaluation standards of the American Evaluation Soci-

2 For an overview of standard sources and their uptake see LL&A et al. (2006, p. 175-176), as well as the European Evaluation Society (http://www.europeanevaluation.org/?page =756983; http://www.europeanevaluation.org/library/evaluation_standards/national_and_ regional_evaluation _societies/europe/index.html)

ety,[3] which have been adopted by, among others, the German Evaluation Society,[4] are important guidelines here for assessing the quality of evaluations with a view to secondary analysis. The four major categories that are defined are **utility** (intended to ensure that an evaluation has served the information needs of intended users), **feasibility** (intended to ensure that an evaluation has been realistic, prudent, diplomatic, and frugal), **propriety** (intended to ensure that an evaluation has been conducted legally, ethically, and with due regard for the welfare of those involved in the evaluation, as well as those affected by its results) and **accuracy** (intended to ensure that an evaluation has revealed and conveyed technically adequate information about the features that determine worth or merit of the programme being evaluated.). These four categories are further differentiated into 30 (US version) and 25 (German version) sub-categories. They are very general, but can serve as a first checklist regarding the quality and soundness of evaluations (see appendix, see also Widmer/ Beyl 2000). The evaluation standards have already been successfully used for case studies in meta-evaluation (e.g. Widmer 1996). Since the general acceptance of the standards is demonstrated by their (sometimes slightly modified) application in various national evaluation societies, and since only some of them are important for our purposes, we can refer to Widmer (1996) for a further discussion of all 30 evaluation criteria. We can concentrate on listing and explaining those that are most relevant to our purposes.

The dimensions **utility** and, above all, **accuracy** are of highest relevance for our purpose of feeding into secondary analysis.[5] Here it is important that an analyst conducting meta-evaluation is capable of assessing the soundness of the various methods used in order to assess reliability and especially the information based on quantitative analysis. In this paper we cannot give checklists of criteria to assess the quality of each and every method used.[6]

A second source for the quality checklist is an important set of standards that has been especially drafted for the field of RTDI policies, i.e. the standards of the Austrian Evaluation Platform (fteval 2003). These guidelines are extremely helpful as they are

---

3 To be found at http://www.wmich.edu/evalctr/jc/

4 See DeGEval; http://www.buero-evaluation.de/DeGeval_standards.htm.

5 For an elaboration of how to understand and use these standards, see the commented guidelines in Joint Committee on Standards for Educational Evaluation (2000): Handbuch der Evaluationsstandards; Opladen.

6 A good overview of evaluation methods that assist in assessing the methodological soundness of evaluations in RTDI policies is presented in Fahrenkrog et al. (2002), Ruegg / Feller (2003) and LL&A et. al. (2006), and, more generally, in the 10 volumes of the CSE Programme Evaluation Kit (http://eric.ed.gov).

targeted at evaluators and policymakers alike in calling for a holist, policy-cycle approach of evaluation. Most importantly, these standards highlight the importance of a sound policy and programme design in the first place as a pre-requisite for sound evaluation.

Thus, in combination with criteria that are being developed within an on-going European project (Edler et al. 2006), the following list of criteria can serve as the basis to assess the quality of evaluations (Exhibit 2). It deliberately limits itself to those criteria that help to decide if an evaluation is good enough to be included in the secondary analysis. To assess the overall quality of the evaluation as part of policy design and implementation, a number of further criteria would have to be checked in the same way (see chapter 4.3.3).

Exhibit 2 also serves as an example of how to score individual criteria. Modifying an approach of Stufflebeam (2000) for the assessment of evaluations, for each criterion an ideal benchmark is defined, against which each evaluation can be scored. These scores then can be fed into the meta-analysis.

Exhibit 2: List and Scoring Table for Basic Criteria to Assess the Quality of Evaluations and their Benchmarks, with a Focus on Accuracy and Appropriateness

Criterion[+]	**Benchmark**	**Score**
Clarity of goals	The goals for the evaluation are derived from the explicit goals of the programme (including their hierarchy and relation) and a clearly and accurately documented evaluation (A1)*.	
Design	The evaluation design – including the mix of qualitative (interviews, case studies) and quantitative methods used – is appropriate, given the objectives of the evaluation and the policy measure.	
Methods[+]	Qualitative and quantitative information are gathered and analysed in an appropriate, systematic way, so that the evaluation questions can be effectively answered. (A7)*.	
Context analysis	The societal, institutional, policy and – if relevant – economic context of the evaluation are examined and analysed in enough detail (A2)*. For technology focused programmes this analysis includes an assessment of the relative position of the technology targeted by the measure vis-à-vis competing or complementary technologies.	
Transparency of evaluation[+]	The purposes, questions, and procedures of an evaluation, including the applied methods, are accurately documented and described, so that they can be identified and assessed (A3)*.	
Quality of information sources[+]	The information sources used in the course of the evaluation are documented in appropriate detail, so that the reliability and adequacy of the information can be assessed (A4)*. All relevant data needed for a certain methodology and to test all programme goals is included.	
Reliability / validity [+]	The data collection procedure is chosen or developed and then applied in a way that ensures the reliability and validity of the data with regard to answering the evaluation questions (A5)*. This includes the usage of transparent indicators for output, outcome and overall success of a programme.	
Systematic data review:	The data collected, analysed, and presented in the course of the evaluation is systematically examined for possible errors (A6)*.	
Clarity of conclusion	The conclusions reached in the evaluation are explicitly justified, so that the audiences can assess them (A8)*.	
Documentation of data and evaluation:	The evaluation should be documented and archived appropriately, so that a meta-evaluation can be undertaken (A9)*.	
Standing of the evaluators	The evaluators are independent and credible and the process of choosing them was transparent.	

Scoring: compared to the benchmark the evaluation is judged to be: 1 completely unsatisfactory, 2 unsatisfactory (room for improvement), 3 satisfactory, 4 above average, 5 best practice.

+ This criterion is a kick-out criterion, any score below 3 leads to the exclusion of the evaluation and the underlying policy measure.

* The code (e.g. A1) means that this criterion is based on or relates to the accuracy criterion 1 of the German Evaluation Society (see appendix).

Source: own compilation, based on the standards of the German Evaluation Societies (focus on "accuracy" standards therein), the standards of the Austrian Evaluation Platform and Edler J., et al. (2006).

The final task of the meta-evaluation in the proposed concept is to check each evaluation if it can be fed into the secondary analysis. What, then, is the selection logic based

on this scoring model? In principle, only if the evaluation meets a certain quality threshold will it be used to characterise the policy measure it assesses. Where exactly this threshold lies and how the scores add up again cannot be defined in a one-size, fits-all standard approach. The decision must be based on a combination of the results of the scoring model with experts' judgement on the overall value of the evaluation. To simplify, each of the dimensions can be weighed in order to take account of the different level of importance for our purposes.

There are, however, some knock-out criteria, such as transparency of the data or independence of the evaluators. Cooksy / Caracelli (2005), for example, report about a meta-evaluation of the evaluations on a agriculture research centre, concluding that the lack of transparency was so severe that a foreseen evaluation synthesis could not be conducted. In Exhibit 2 these knock-out criteria are labelled with a ($^{+}$). These criteria must not only be weighed high compared to other criteria, but without a score of at least 3 (on a scale between 1 and 5) in this criterion, the evaluation should not be used for the secondary analysis. For some other criteria, the results of an evaluation can still be valuable for the secondary analysis, even if the score is below 3. Examples here would be reporting standards, which might not meet average standard but still allow for an interpretation of results.

2.4 Characterisation of Policy Measures

After the evaluations have been classified and checked for quality, the underlying policy measures for which the evaluations are sound and useable must also be characterised and described in more detail and coded in a data sheet.

This is indispensable in order to use the measure for the meta-analysis and for the evaluation synthesis. For the data sheet and further analysis, the formal quality of the policy measure in terms of its design, management, aims, policy context etc. needs to be assessed and the performance of the policy measure in terms of impact, effectiveness and efficiency of the policy measure have to be assessed.

In principle, the policy measures can be classified in two dimensions (1) basic characterisation and (2) quality (performance). As with the evaluations, for each of the categories within these two dimensions below simple numerical codes or assessment scores (in a scorecard approach) can be applied, combining quantitative and qualitative data, and enabling comparison and aggregation.

For the purposes of the secondary analysis of policy measures and policy mixes, the crucial step is to qualify the performance of a measure. This will be done on the basis of the evaluations that passed the quality check (see above). Those which did not pass

this test or for which no evaluation was conducted cannot be included in the database and in the secondary analysis to the same degree. For the secondary analysis, however, the knowledge of existing policy measures, even if not based on a sound evaluation, is indispensable. Thus, for the sake of the evaluation synthesis all relevant policy measures should be included in the analysis as far as possible at the level of description.

2.4.1 Basic Characteristics of the Policy Measure

Indicative dimensions used to characterise the nature of a policy measure include:

- Defining the target function: this is the most important element of the characterisation, as the target function is the benchmark for the evaluation and performance of the programme. Both the overall objective and the concrete measurable targets have to be included in this characterisation. If defined, the measures with which success will be ascertained later on need to be included as well as the concrete benchmark for each of the targets. Potential dimensions for traditional target functions in R&D programmes are input additionality, output additionality and behavioural additionality for the beneficiaries, as well as systems impact like improved innovative dynamics and competitiveness.
- Policy level: European, national, regional.
- Geographical reach: availability to regional actors, actors nationwide or international actors.
- Type of policy measure: the typology of policy measures follows from the objective of the analysis. To assist the selection, one can follow established categorisations (see for example Exhibit 3). Other categorisations can be found in Georghiou et al. (2003), Jochem et al. (2006) or for innovation policies in a narrow sense, in the TrendChart database of the European Commission.[7]

7 See trendchart.cordis.lu/

Exhibit 3: A Categorisation of Measures in RTDI Policy

Finance	*Educate, raise awareness*
1. *Institutional support of public research institutes* - universities - non-university institutes 2. *Financial incentives for research especially in industry* - direct financial incentives for research and experimental development (support programmes) - tax exemptions	5. *Education and formation* - university education - technical colleges - on the job training 6. *Innovation management* - support of organisational adjustments in industry (and public institutes) - support of absorptive capacities - innovation consultancy 7. *Awareness building and scientific consultancy* - technology assessment and forecast - targeting, long-term vision building - scientific advisory committees
Structure, integrate	*Regulate and spur economic activity*
3. *Infrastructure* - technology parks - support of technology clusters 4. *Measures to increase competition and technology transfer* - programmes for vertical and horizontal co-operation - networks, competence centres - intermediaries	8. *Regulative measures* - regulatory policy (IPR, , standards, norms, labour regulations etc.) - competition policy - innovation-friendly büreaucratic and regulative framework 9. *Creation of companies* - risk capital - start-ups 10. *Demand-oriented policy* - public demand - spur private demand (financial incentives, enabling)

Based on Meyer-Kramer / Kuntze (1992), own extension and modification

- Responsible institution(s): the institution responsible for the design and implementation of the measure has to be defined for each measure. Here, it is especially important to differentiate between the political responsibility in ministries and the operational implementation in agencies. This distinction differs between countries and ministries.
- Scope of instrument: distinction between a single programme, a set of programmes and measures or if the object of analysis even can be defined as a "policy". Which level of the measure is to be analysed is of crucial importance. For example, there are many multi-measure, multi-actor programmes (MAP) established, mainly in the form of broad competence centre programmes.[8] For a secondary analysis one would have to define at which level these MAPs are to be analysed.
- Interplay of instruments and policy context: policy measures with which the analysed measure interacts, e.g. a pre-existing or complementary scheme. How deliberately is this done and to what extent is the policy measure designed to complement oth-

8 For an overview of such programmes, see FFG et al. 2004, Appendix.

ers? What is the policy context and institutional environment of the measured policy?

- Thematic scope: a characterisation of measures would also differentiate between thematic programmes on the one hand and horizontal measures on the other hand.
- Thematic programmes can be further differentiated into sub-programmes (specific technologies) and actor groups targeted. Does the measure explicitly or implicitly target a clearly defined technological or knowledge area? Is it a programme to support a specific industrial sector?
- Principles of design process sand governance: how is the programme design organised, what are the information sources and how is interaction with stakeholders, if any, organised? How transparent and accessible is the policy design process to outsiders? Can strategic and operational responsibilities clearly be attributed to specific actors? Are there feedback loops within the policy or measure cycle?
- Functional mechanism and incentives: type of measure and incentives, e.g. direct grants, loans, mobility incentives, infrastructural support, information provision, training, advice.
- Weight of incentives: percentage of the funding quota, relative importance of the scheme for the target group.
- Behavioural conditions for support: here all conditions have to be mentioned which are pre-conditions for funding within a programme. For example, is multi-actor cooperation a prerequisite, if yes, in what form? Is increased R&D input asked for, if yes, to what extent?
- Definition of target groups: for example, firms in general or only SMEs, research organisations, HEIs, individual researchers, also inclusion of international actors.

2.4.2 Quality of the Policy Measure

Compared to the assessment of evaluations, there are even less standardised approaches to assess the quality of a policy measure. It is self-evident that the judgement of performance of a measure needs to be derived directly from the target function and type of the underlying measure. Thus, ex ante only general categories can be defined that have to be tailormade for each concrete analysis. All information that feeds into this assessment should be based on sound evaluations as defined above. In general, the better the evaluation of a policy as assessed above, the more differentiated and detailed a qualification of the policy measure can be.

- overall impact, including impacts on various target groups
- added value (in terms of all relevant dimensions: input, output, behavioural)
- goal attainment and effectiveness
- implementation and cost-efficiency

- design: appropriateness given the context of the measure (policy mix, policy gaps, problem definition)
- management
- role of evaluation in the whole policy-cycle[9] (see category "role in policy cycle" in the assessment of evaluations
- originality / novelty / creativity of the policy measures.

For the policy measures where no evaluation exists or the evaluation does not meet the threshold of quality, only a short characterisation based on available programme documents can be done. These descriptions need to be used for the evaluation synthesis later on as additional context information, but they cannot be used for the in-depth secondary analysis (meta-analysis), as important information on quality and effects is missing.

3 A Deeper Understanding at the Instrument Level: A Meta-analytical Approach

3.1 Definition and Purpose

The term meta-analysis refers to the statistical and quantitative analysis of a large collection of analyses, which document the results of individual studies (Glass 1976, p. 3). Meta-analysis is a collection of conceptual and methodological approaches to summarise the empirical evidence for a given research question (Beelmann & Biesner 1994, p. 211). It is a method following the paradigm of empirical research to achieve the quantitative integration of the results of various empirical studies and to shed light on the variability of these results (Drinkmann 1990, p. 11). Meta-analysis serves as a systematic and quantitative alternative to narrative literature reviews and may have advantages for summarising the effects of policy measures reported in single evaluations if the number of evaluations is too large to oversee.

Meta-analysis is an established method and comprises an established set of tools dating back to the first systematic and quantitative reviews in medical journals in the 1950s to the prominent discussions in economics e.g. in the *Journal of Economic Perspectives*. However, it is argued here that meta-analysis is grossly underutilised in discussing the effectiveness of innovation policy measures and related policy learning.

---

9 This category mirrors the category of "role within the measure cycle" for the evaluations outlined in chapter 2.2.1, for further details see there.

Meta-analysis is a distinct step in our secondary analysis which is not to be confused with meta-evaluation. Essentially, meta-evaluation aims at evaluation and assessing the quality and the type of evaluations and the covered policy measures. Meta-analysis, however, is not concerned with analysing the quality of evaluation studies. Rather, it focuses on three dimensions which are not covered by meta-evaluation. These dimensions define three different types of meta-analyses: analysis of the effects of policy measures, examination of the effect of control variables and examination of new hypotheses (Miller and Pollock 1994).

- **Type A: Analysis of the Effects of Policy Measures**

  Type A meta-analyses investigate the size of effects of the policy measures. The integration of several individual studies increases the overall number of observations, which are used to estimate the effects. These higher number of observations result in better estimates of the effects. The sole aim of a Type A analysis is to increase the accuracy of the estimate of an effect more than would be possible in a single evaluation. A typical research question would be: Does participation in a national funding programme have an effect on the generation of innovation? By integrating more evaluation studies in a meta-analysis, one would investigate the participation indicator variable.

- **Type B: Examination of the Effect of Control Variables**

  Type B meta-analyses are not primarily concerned with the effect of the policy measures, but with the impact other control-variables have on the effect of the policy measure. These control variables essentially are part of the individual evaluations, but are not considered in the course of the evaluative discussion. Type B meta-analyses thus focus on explaining the variability of the results of the individual evaluations. In this context, a research question would be: Does the policy measure generate different effects with small and medium-sized enterprises than it does for large companies? By integrating more evaluation studies by means of a meta-analysis, one could investigate the indicator for the fraction of SMEs in the analysed data set, which was controlled for in the individual evaluations. Although the analysis of a control variable does not have an immediate bearing on the assessment of the policy measure, it can have strong implications on the targeting and the overall effectiveness of the policy measure and its impact in the broader set of policy measures.

  For illustration: consider a set of policy measures (funding programmes) which are targeted towards alleviating the financial constraints commonly faced by innovating SMEs. Each evaluation of an individual policy measure may reveal that the participating companies have a higher likelihood of commercialising new products than

non-participating companies. Hence each programme can be associated with a strong effect. Additionally, imagine that each individual evaluation shows that there is no difference in the probability to produce product innovations between SMEs and large enterprises. One would conclude that each policy measure has successfully targeted the financial constraints and enabled SMEs to be as innovative as larger companies. Yet, a meta-analysis where the results of all the individual evaluations are pooled can show that the indicator variable for the fraction of SME in the sample of the individual evaluation still has a negative significant effect on the effect of the policy measure. In this case the meta-analysis would integrate the results of all individual evaluation studies, the conclusions, however, would be the opposite. The policy measures are not successful in fully removing the unfavourable conditions for SMEs as – even with accounting for programme participation – SMEs show a systematically lower likelihood to innovate.

- **Type C: Examination of New Hypotheses**

  Type C meta-analyses move beyond the analysis of control or moderator variables as in meta-analysis of type B. Meta-analyses of type C utilise variables generated from the information in the primary studies which can be analysed once various evaluation studies are available. The variables which are under consideration here cannot be tested in the primary evaluations. This can be due to the fact that across the individual primary evaluation this variable does not change. An example for a type C meta-analysis would be to investigate whether self-evaluations and evaluations carried out by independent evaluators report different magnitudes of effects.

3.2 Process of the Meta-analysis

An ideal type meta-analysis proceeds in four steps: (1) elaboration of research question, (2) collection of relevant documents, (3) coding and assessing the evaluation studies and (4) data analysis. In our proposed concept of a secondary analysis, the first three steps have already been taken in the process of meta-evaluating the individual evaluations. Thus, the following discussion contains supplementary information on the coding step and then concentrates on data analysis.

3.2.1 Coding of the Information

The coding of the information in the evaluation studies is the key to a valid and interesting meta-analysis. Based on the research question and taking account of the different types of evaluation studies at hand, a code book of the information required for the analysis has to be devised. It has to define the setup of the data base, either as a flat file for simple meta-analysis tasks or relational data bases for more complex tasks. A

data base infrastructure is particularly recommended where most evaluations supply information on more than one indicator of the effect.

3.2.1.1 Effect of the Policy Measure

The coding of the effects differs between the situations where all evaluations in the meta-analysis supply quantitative information and where some of them supply quantitative and others only qualitative assessment of the effects.

Quantitative information

If only quantitative evaluations are available for the meta-analysis, the effect of the policy measure has to be retrieved from the study and coded in the data set. Generally, two indicators are suggested in the literature as a measure of the effect: the mean difference and the correlation. As discussed in DeCoster (2004), these two indicators can be derived from a host of statistical information given in quantitative studies. The mean difference (Cohen's g) can be correctly derived from (1) between subject test statistics such as t-statistics, z-statistics, one-way ANOVA, F-statistics, (2) indirect calculation based on two-way ANOVA results, (3) within subject test statistics (4) from p-values (5) dichotomous dependent variables, (6) averaging other effects and (7) correlation coefficients. So even if the statistical approaches in the evaluations under investigation are not the same, the estimated effects can be transformed so that it is consistent for the different observations (evaluations).

Quantitative and qualitative information

If quantitative and qualitative evaluation studies are available, it is our understanding that meta-analysis should be able to integrate also qualitative and quantitative evaluations in the quantitative assessment of the effects. Both the results of qualitative and quantitative studies can be interpreted to show effects on a say five-level Likert scale ranging from *strong negative* to *strong positive* effects. While coding the evaluation studies, the coder has to interpret the results of the quantitative and the qualitative studies. In this case, intercoder reliability has to be established. In addition to the ordinal coding of the effect, collecting the data would also imply specifying what type of evaluation the effect is derived from (qualitative/ quantitative) and what kind of effect was measured. First coding quantitative information on the effects as suggested above and second coding both quantitative and qualitative information on the effects into the ordinal scale supplies standard data for the quantitative studies. It can be utilised in a meta-analysis on the sub-sample of quantitative evaluations.

3.2.2 Analysing the Data

In this section, the actual analytical step in the in meta-analysis is discussed by means of an illustrative example. This example does not show all analytical techniques. Rather, it tries to point to some of the analytical techniques without discussing them in full length.

Table 1 contains the data for an illustrative example.[10] It contains a compilation of the artificial results of 30 evaluations of funding programmes for companies. From each of the evaluations a set of information was retrieved – a discussion of the potential data which can be collected from given evaluations can be found below – and compiled in a single table. Both for the funded companies and a control group of non-funded companies, the number of innovations x which were commercialised by these firms was retrieved. The innovation rates can be calculated from this information. The ratio of innovation rates (RR) is greater than one, if the rate of innovations in the funded group is higher than in the control group. It is a measure of the effect of the policy measures. As theory suggests the natural logarithm of the rate-ratio is more normally distributed, one would continue with the natural log of the ratio of innovation rates. It is denoted Ln(RR). The standard error SE(Ln(RR)) of the Ln(RR) is the square root of the inverse sum of innovations of founded companies and the control group. Given the standard error of Ln(RR), we can compute the 95% confidence interval for the Ln(RR) and from that the 95% confidence interval of RR. These are graphed in Table 2. Where the whole confidence interval is larger than 1, the evaluation suggests a significantly positive effect of the policy measure. Evaluation No. 1 hence points to a successful policy measure. Evaluation No. 4 suggests that the overall effect of the programme is negative. The innovation rate among the funded firms is smaller than the rate of innovation among the control group, giving rise to an estimate of the effect which is significantly smaller than one. Evaluation No. 3, however, suggests that the hypothesis that the estimation of the effect equals one cannot be rejected based on the finding, as 1 is contained in the confidence interval. For the whole approach, see Hedges et al. (1999).

A measure of the pooled effect – Type A meta-analysis

A measure for the pooled effect of the policy measures is the weighted average of the effects, where the inverse variance of the individual effect measure is used as a weight (Fleiss 1993). The pooled variance of the log of the effect is the inverse sum of the weights. Based on the measure of the pooled effect and its variance, the confidence intervals for the pooled effect can be determined. It is also graphed in the bottom part

10 The example is inspired by Sutton, Abrams and Jones (2003).

of Table 2 and labelled appropriately. The pooled size of the effect is 1.11 which is greater than 1 at the given level of significance, suggesting positive effects of the programme on the innovation output. In addition, one clearly observes that the pooled effect has a considerably smaller confidence region, indicating an improved accuracy of the effect estimate.

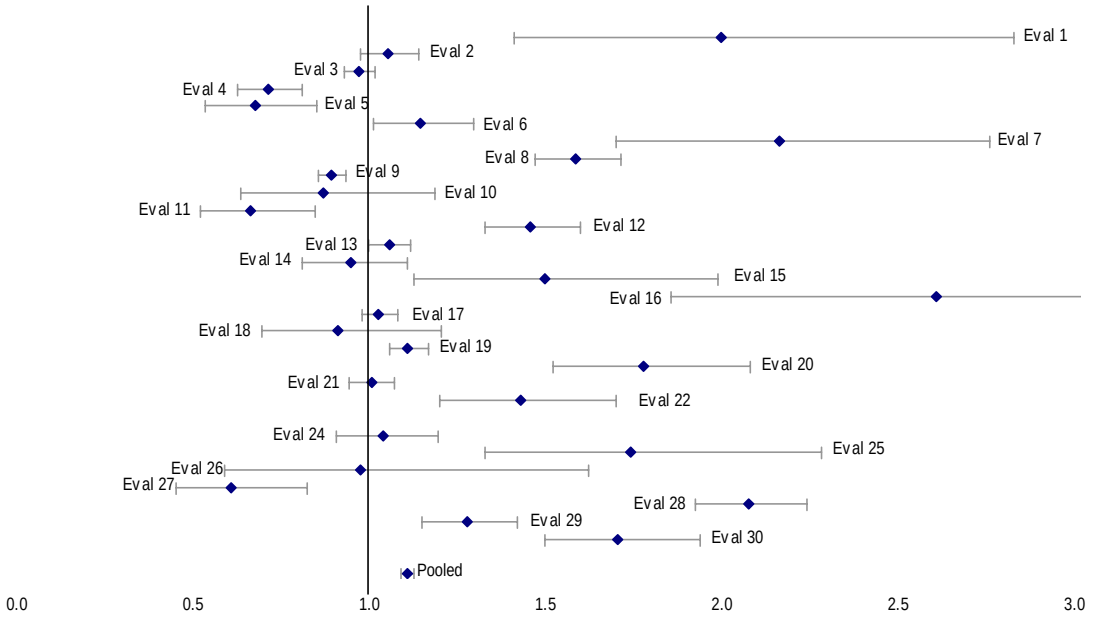
The underlying assumption of the pooled estimate here is that the underlying evaluations measure the same treatment effect. Yet it can be argued that the layout of the policy measures, the setup of the evaluations etc. differ across the studies and influence the results of the evaluations. The evaluations are not identical replications of one another. The pooled estimate is expected to vary by more than just by chance as would be the case if the evaluations were identical replications. In contrast to the fixed effect approach here, random effects can be integrated in the pooling of the evaluation results and account for the fact that the effects measured are likely to be drawn from a distribution of effects rather than being random variations of identical effects. The random effects models account for the variability of the effects, however, they do not explain it.

Table 1 Illustrative Example

		Funded companies			Control group								
	years	No. of firms	No. of innov..	Innov rate	No. of firms	No. of innov.	Innov rate	R.-ratio (RR)	Ln(RR)	SE (Ln(RR))	Weight	Self eval.	Fract. of SME
Eval 1	1.5	444	48	0.072	444	96	0.144	2.000	0.693	0.177	32.000	0.000	0.115
Eval 2	3	5568	1224	0.073	5268	1224	0.077	1.057	0.055	0.040	612.000	0.000	0.155
Eval 3	6	5988	4164	0.116	6024	4080	0.113	0.974	-0.026	0.022	2060.786	0.000	0.302
Eval 4	6	1376	576	0.070	1392	416	0.050	0.714	-0.337	0.064	241.548	0.000	0.298
Eval 5	12	1140	180	0.013	1120	120	0.009	0.679	-0.388	0.118	72.000	1.000	0.261
Eval 6	2	624	464	0.372	656	560	0.427	1.148	0.138	0.063	253.750	0.000	0.114
Eval 7	3	320	96	0.100	320	208	0.217	2.167	0.773	0.123	65.684	0.000	0.218
Eval 8	3	284	1108	1.300	280	1736	2.067	1.589	0.463	0.038	676.332	1.000	0.098
Eval 9	6	9730	4802	0.082	9814	4340	0.074	0.896	-0.110	0.021	2279.663	1.000	0.334
Eval 10	2	356	86	0.121	352	74	0.105	0.870	-0.139	0.159	39.775	0.000	0.289
Eval 11	6	540	162	0.050	540	108	0.033	0.667	-0.405	0.124	64.800	1.000	0.288
Eval 12	6	1728	756	0.073	1746	1116	0.107	1.461	0.379	0.047	450.692	1.000	0.064
Eval 13	3	6666	2288	0.114	6600	2398	0.121	1.059	0.057	0.029	1170.854	1.000	0.246
Eval 14	6	1200	408	0.057	792	256	0.054	0.951	-0.051	0.080	157.301	1.000	0.303
Eval 15	1	400	80	0.200	400	120	0.300	1.500	0.405	0.144	48.000	1.000	0.118
Eval 16	1.5	348	48	0.092	300	108	0.240	2.610	0.959	0.173	33.231	0.000	0.199
Eval 17	12	2664	3168	0.099	2680	3280	0.102	1.029	0.029	0.025	1611.514	0.000	0.088
Eval 18	3	840	108	0.043	816	96	0.039	0.915	-0.089	0.140	50.824	0.000	0.327
Eval 19	9	2508	2970	0.132	2502	3294	0.146	1.112	0.106	0.025	1561.810	0.000	0.107
Eval 20	6	744	252	0.056	696	420	0.101	1.782	0.578	0.080	157.500	1.000	0.150
Eval 21	12	3582	1926	0.045	3690	1998	0.045	1.007	0.007	0.032	980.670	1.000	0.084
Eval 22	12	630	220	0.029	600	300	0.042	1.432	0.359	0.089	126.923	1.000	0.165
Eval 23	6	280	80	0.048	320	408	0.213	4.463	1.496	0.122	66.885	0.000	0.076

Eval 24	6	816	392	0.080	816	408	0.083	1.041	0.040	0.071	199.920	0.000	0.240
Eval 25	6	560	96	0.029	388	116	0.050	1.744	0.556	0.138	52.528	1.000	0.210
Eval 26	3	270	30	0.037	276	30	0.036	0.978	-0.022	0.258	15.000	1.000	0.312
Eval 27	4	490	110	0.056	510	70	0.034	0.611	-0.492	0.153	42.778	0.000	0.302
Eval 28	6	3540	980	0.046	3720	2140	0.096	2.078	0.731	0.039	672.179	0.000	0.145
Eval 29	3	1524	616	0.135	1524	788	0.172	1.279	0.246	0.054	345.732	1.000	0.203
Eval 30	2	1536	352	0.115	1760	688	0.195	1.706	0.534	0.066	232.862	0.000	0.217

Table 2: Confidence Intervals of Individual and Pooled Estimates of the Effects



Explaining heterogeneous effects – Type B and Type C meta-analyses

One possibility to explain heterogeneous effects is to use meta-regression. This regression technique does not use primary data to assess the effect of a policy measure or the effects of a set of policy measures. It is based on the data collected by the meta-analyst, where each observation in the regression data set is the result of an individual evaluation. For illustration we employ a meta-regression on the data given in Table 1.

In this regression, we want to investigate whether the size of the firm has an impact on the effect of the policy measure. For this purpose Table 1 contains information on the fraction of SMEs in the group of funded companies. A weighted OLS regression of Ln (RR) on the fraction of SMEs and the constant is reported in Table 3. It shows that the overall effect declines with the participation of SMEs. Where there are no SMEs in the funded firms, the overall effect of the policy measure would be 1.48. With 30% of SMEs among the funded firms, the overall effect shrinks to 0.937 which is below 1 and hence indicates a negative impact on the average.

Table 3 Meta-regression I

```
      Source |       SS       df       MS              Number of obs =      30
-------------+------------------------------           F(  1,    28) =   10.55
       Model |  .545787832     1  .545787832           Prob > F      =  0.0030
    Residual |  1.44903532    28  .051751261           R-squared     =  0.2736
-------------+------------------------------           Adj R-squared =  0.2477
       Total |  1.99482315    29  .068787005           Root MSE      =  .22749

------------------------------------------------------------------------------
        lnrr |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sme |  -1.524345   .4693879    -3.25   0.003    -2.485843   -.5628477
       _cons |    .392197   .0987011     3.97   0.000     .1900169    .5943771
------------------------------------------------------------------------------
```

In Table 4 the indicator for the type of evaluator is included in the regression to test whether self-evaluations report different magnitudes of the effects than professional and independent evaluators do.

Table 4 Meta-regression II

```
      Source |       SS       df       MS              Number of obs =      30
-------------+------------------------------           F(  2,    27) =    5.09
       Model |  .545799378     2  .272899689           Prob > F      =  0.0134
    Residual |  1.44902377    27  .053667547           R-squared     =  0.2736
-------------+------------------------------           Adj R-squared =  0.2198
       Total |  1.99482315    29  .068787005           Root MSE      =  .23166

------------------------------------------------------------------------------
        lnrr |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sme |  -1.525515   .4846076    -3.15   0.004    -2.519848   -.5311825
    selfeval |   .0012622   .0860508     0.01   0.988    -.1752996    .1778239
       _cons |   .3918409   .1034032     3.79   0.001      .179675    .6040068
------------------------------------------------------------------------------
```

The type of evaluation, either a self-evaluation or an evaluation carried out by an independent and professional evaluator, does not influence the effect of the policy measure in this artificial data set.

3.3 Observations on the usage of meta-analyses

Meta-analysis is a powerful tool for integrating characteristics of the studied policy measure, its effects, design and features into a single analysis. However, several aspects also have to be taken into account when conducting a meta-analysis. A common

objection against meta-analysis is that it compares apples and oranges. However, the interesting part in the meta-analysis as we use it in the secondary analysis of evaluation studies is to analyse how the setup of policy measures determines their effect. It is only a marginal issue how the evaluation setup determines the reported effects. Meta-regression controls for the differences in the study design and other characteristics to validate the analysis.

A second objection against the usage of meta-analysis is the problem of “garbage in – garbage out”. Quality issues are at the heart of the argument that inferior evaluations will generate inferior results in the meta-analysis. As discussed above, the meta-evaluation step takes care that no qualitatively inferior evaluations are considered in the subsequent steps. Moreover, as seen in the illustrative example, meta-analysis is capable of increasing the accuracy of the estimate of the effect. Garbage in the individual studies in terms of measurement error is reduced by the meta-analysis.

Another aspect that has to be taken into account when doing a meta-analysis is the possible selection bias of the underlying evaluations: if evaluations are used for accountability reasons in the political process, we can expect that evaluations are more likely to be commissioned for programmes which are supposed to be successful. A further sub-class of the selection bias is the publication bias. Broad search strategies including the search for un-published material can be rather expensive, both in terms of search cost and time spent tailoring the search strategies. In addition, Rosenthan (1979) and Oswald (1983) suggest the fail-safe-N as a measure for the validity of the results where publication bias is suspected. Essentially, it answers the question about how many other (and negative) studies one would need to cover through the results of the meta-analysis. Large N suggests that – even if publication bias exists – it has a rather small impact on the validity of the results.

For the evaluation of policy measures, meta-analysis does not only offer an opportunity to pool given evaluation studies and compute new results and interesting insights from the primary results of the individual evaluations; meta-analysis also offers the possibility to overcome data access constraints which for example limit the possibility to asses the overall effects of European Framework Programmes on companies. A perfect data source for this task is the Community Innovation Survey, which is carried out in all 25 Member States, Norway, Iceland, and the candidate countries. Pooling these data sets is not possible as national data confidentiality policies restrict access to the datasets to the national statistical offices. A research programme anticipating meta-analytic final steps can overcome these constraints and carry out an overall evaluation of the effectiveness of the framework programmes. National evaluations – based on the same

data set and employing a comparable set of techniques – can be aggregated into a pooled measure of the effectiveness.

4 Understanding Policy Performance in a Systems World. Applying an Evaluation Synthesis

4.1 Definition and purpose

The concept of evaluation synthesis is less clearly defined than that of meta-analysis. Evaluation synthesis is a secondary analysis to answer questions on a set of evaluated policy measures that cannot or only at high costs be answered by a single evaluation. Evaluation syntheses take advantage of existing studies and interpret their findings in a new way. This enables the interplay and complementarity of measures and policies to be analysed and a new quality of knowledge based on existing knowledge to be derived.

The modified evaluation synthesis we apply is based on an evaluation synthesis approach that goes back to Light (1984) and which has been further developed and intensively used by the United States General Accounting Office (GOA 1992, see also Scriven 1991, Borgmann 2005, Cooksy/Caracelli 2005). Evaluation synthesis has been defined as a "combination of the results from more than one study in order to come to general statements about an intervention" (Cooksy/Caracelli 2005, p. 32, following GOA 1992), and similarly, as a "content synthesis of multiple evaluation reports on similar programmes or projects" (Beywl & Associates 2004).

The evaluation synthesis approach has mainly been used to assess the benefit of one specific support programme or a medical treatment by combining the insights of a (mostly high) number of evaluations. We propose to take advantage of many of the principles of such an evaluation synthesis approach as developed in the evaluation literature. However, in our approach we modify this understanding of evaluation synthesis according to the needs and limits in the field of RTDI policy. In RTDI policy we need a system analysis (see chapter 4.2) that can benefit from a secondary analysis of existing evaluations. At the same time, in contrast, for example, to clinical research in RTDI policy we rarely find a high number of evaluations for any given individual programme or intervention, thus making evaluation synthesis for one specific intervention impossible.

The major purpose of the evaluation synthesis in our approach derives from the application of the evaluation synthesis that is based on evaluations not of one single programme, but on a number of programmes in a given innovation system (or for a certain

target group). This purpose is to support the system perspective on policymaking. An evaluation synthesis should enable the identification of gaps, redundancies, re-enforcing effects (complementarities) and thus the performance of policy mixes in a given innovation system. In addition, in combination with the systematic preparation of such an evaluation synthesis as laid out in chapter 2.3, it should enable qualified assessments of the evaluation culture and capabilities in a given system.

Since the different programmes and measures that have to be included differ in their target function and design, and since the number of measures tend to be much more limited as compared to meta-evaluation, the modified evaluation synthesis will largely be done in a qualitative manner. Thus, while the meta-analysis is a quantitative approach that deals with better understanding of individual measures, the (modified) evaluation synthesis is a largely qualitative approach to better understand the nature of policy mixes at a systems level.

To highlight the value of evaluation synthesis for policy understanding in the systems world, a short chapter on the specific need for such an approach in the field of RTDI policy in the first place and on the reasons why one should take on the costs of such a complex exercise, precedes the description of the evaluation synthesis approach.

4.2 The Need for Evaluation Synthesis in RTDI Policymaking

Evaluation synthesis as defined here is of special relevance for RTDI policymaking, due to the innovation system world in which RTDI policymaking operates – which it also creates - and due to the overall complexity of cause and effects in this area. RTDI policies tackle actors who are embedded in a system of actors and institutions, and they do not function in isolation, but within a network of other policy measures. The innovation systems approach – in all its various shapes describing national (Nelson 1993; Lundvall 1992, Edquist 1997), sectoral (Malerba 2002, 2004), regional (Howells 1999, Cooke et al. 1998) or technological (Carlsson and Stankiewicz 1991, Carlsson 1995, 1997, in particular 2002) innovation systems – is a heuristic that highlights the importance of the interplay of actors and actor groups, institutions and policy measures. In this systemic approach based on evolutionary economics, the legitimation for policy lies in a set of system failures, such as capability failures, failures in institutions, network failures and framework failures (Arnold 2002). Arnold (2002) summarises the challenges of policy-making in a systems world and the consequences this has for evaluation, and he even provides a set of criteria for the evaluation of innovation systems instead of individual innovation policies. Consequently, intelligent policymaking – both when designing individual measures and a mix of policies – must take this inter-

play and the combined effects of policies into consideration. To do so, adequate strategic or distributed intelligence (Kuhlmann et al. 1999; Kuhlmann 2001; Georghiou 1995, Georghiou / Roessner 2000) is needed in order to base decisions on the knowledge about effects on the systems level and the interplay of a given measure with other measures and institutional framework conditions in the system.

We argue that one way to provide more adequate strategic intelligence in a systems world can be based on the systematic usage of existing evaluations of individual measures by means of an evaluation synthesis. The alternative to this approach would be a comprehensive analysis at the system level. One such an example is the evaluation of the Norwegian Research Council (Arnold et al. 2001), in which the evaluation of the portfolio of one large organisation resembles a system analysis because of the specific portfolio in Norway. Another example is the comprehensive evaluation of the Basic Plan in Japan (Blanpied 2005, Kondo 2005). However, these comprehensive approaches are costly and they raise a number of institutional challenges due to the co-ordination needed, as in most innovation systems a multitude of actors is responsible for the various policy measures – rather than one key ministry or council. In consequence, although many policymakers nowadays acknowledge the need for evaluation of policy mixes at the systems level, such approaches are the exception rather than the rule.

For the system question a qualitative evaluation synthesis is useful for two most obvious reasons:

1) most existing evaluations focus on a specific measure situated in a particular context, .

3) the effects of programmes are assessed on different levels and for different target groups. Simple calculation of aggregated effects makes no sense, and expert judgement as to the relative weight of levels and programmes is crucial.

4.3 The Process of an Evaluation Synthesis

4.3.1 Options and starting points

There are various options for using evaluation synthesis. The first option is to conduct an evaluation synthesis for one specific intervention. This would in fact be the model that resembles the origin and most comprehensive application of evaluation synthesis (GAO 1992). However, the pre-condition for such an analysis would be that a number of evaluations exist for one measure. In RTDI policy, in contrast to, for example, medical treatments, the likelihood of having a sufficient number of evaluations for such an

analysis is generally low. One – rare – example is the US Advanced Technology Program that has been analysed in depth using various methods and with different objectives studies (Ruegg/Feller 2003, chapter 9; Ruegg 2005). In that case, a further selection of policy measures is in principle not needed. However, even such a one-measure synthesis can benefit from the review of those programmes that influence the ATP, e.g. other technology-oriented support programmes at the level of the states in the USA.

This leads us to the second, more realistic and likely option, which is to consider the overall effects of a mix of programmes or a framework type of programme approach on a certain sub-system (technological, sectoral, national, regional) and/or an actor group. The analytical perspective and the questions to be asked in this approach are manifold:

- The broadest possible approach would be to ask if the policy mix in an innovation system properly assists the functions innovation systems should perform. For such a basic approach we can draw upon a large set of compilations of innovation system functions (among others Edler et al. 2006b; Hekkert et al. 2006; Borrás 2004). Such an evaluation synthesis would be targeted at the "highest hierarchies" and/or for the programme mangers responsible for individual programmes, in order to learn the relative contribution of their programme to the system's performance. A full range synthesis at highest level, however, would be a very challenging and comprehensive endeavour. Therefore, such analysis would be more realistic when confined to a regional, technological or sectoral sub-system.
- In a more limited scope, all programmes/measures that contribute to fulfilling one concrete function in a system, such as, for example, assisting financing of start-ups could be examined.
- Even more limited, but a further typical question for an evaluation synthesis could be to examine the overall effect and interplay of a clearly defined set of measures specifically targeting a clearly defined actor group.
- In a more focused approach that is interested in the system thinking of policy-makers, one can strive to analyse and compare the rationales and the goal perspective of the evaluated programmes and institutions: are these derived from an overall strategy and does this all fit together in a strategy?
- The key for an analysis of a set of policies is to ask for the relationship of measures, i.e. to analyse gaps, bottlenecks, complementarities, contradictions and unnecessary redundancies in the context of the innovation system and its failures. Such an analysis has to include system level issues of efficiency and effectiveness: how can programmes be thought and brought together in order to better achieve policy aims?
- What lessons can be learned via the comparison of programmes evaluated: management failures and good practice?

Our concept concentrates on the evaluation synthesis approach that analyses a set of measures rather than one measure. The selection of policy measures is crucial in this approach. It follows strictly from the research question and the system levels defined and requires a broad screening process. Ideally, such a modified, multi-measure synthesis covers all policies that affect the actors of a given system regarding the target

function. Thus, a comprehensive evaluation synthesis would have to include not only the measures that are implemented by the political institutions in a given system, but also at the neighbouring levels as these programmes affect the performance of the sub-system being researched. Furthermore, such an analysis would have to ask which institutions are responsible for policies that impinge upon the development and market diffusion of a technology.

One example for such an approach would be to understand the role of RTDI policy for the advancement of fuel cells in Germany. To understand the role of policy, one would not only include evaluations of programmes of the ministry mainly responsible for energy research (BMWi), but also programmes of the federal environment (BMU), transport (BVBS) and research ministry (BMBF) , which all have responsibility for this technology. In addition, the evaluations of the EU Framework Programme as well as regional fuel cell initiatives in the German federal states must also be screened. The evaluation synthesis characterises all relevant measures and then asks how all these policies work together, how their targets, incentives and effects inter-relate, what the specific role of a certain level or institution is in this complex web of policies and how target groups are affected by individual interventions and by the policy mix.

One concrete example of such an approach was the study on RTDI policies in Germany that are geared towards improving the technological competitiveness of the eastern German federal states (Koschatzky/Lo 2005). In this study, the relevant policies of federal ministries and state ministries, as well as the role of the European support programmes were systematically screened and reviewed. A further area in which a range of evaluations was reviewed is the assessment of the European Framework Programme (Vonortas/Hinze 2005; Arnold et al. 2005). The data availability for a full-scale evaluation synthesis, however, is insufficient.

4.3.2 The synthesis process

Characterisation of evaluations and policy measures - building upon the meta-evaluation

For all relevant policy measures, one needs to collect and select the evaluations done. The cornerstone of the subsequent data interpretation in the evaluation synthesis is the meta-evaluation (see chapter 2.3). This defines the characterisation and assessment of the evaluations and lays the basis for the analysis of policy measures, as once an evaluation is rated sound and appropriate, its results can be fully used, systemised and interpreted to characterise the selected policy measures (chapter 2.4). This is the cornerstone of the cross-cutting analysis. It is important to note that also those policies

shall be considered for which a sound evaluation has not (yet) been conducted. Those policy measures still need to be characterised as far as possible on the basis of descriptive information available, as they are an important interdependent variable in assessing the overall performance of a policy mix.

The relation of evaluation synthesis target and measure targets

An evaluation synthesis is a narration of policy implications with the theme defined by the target of the evaluation synthesis – not by the targets of the individual measures or the targets of the evaluations used. Thus, all the targets of the policies need to be related not only to each other, but also to the evaluation synthesis target. This means that some of the targets of the measures might be less relevant, while others form the core of the analysis. The targets of the measures can be weighed according to their relevance for the evaluation synthesis goals. On the basis of the evaluations, the degree of goal attainment for all the relevant measures and the relevant targets can be compiled and the relative contribution of individual policies can be assessed. In complex evaluation synthesis the overall target may be better split into sub-targets, and the target contributions of the individual measures can be attributed to those sub-targets. Ways of depicting these relations can be a target tree or a target/measure matrix. Such a presentation will already make assessments of the inter-relation (complementarity, contradictions, doubling) of targets.

Furthermore, the relation between the various targets in different instruments (vertically and horizontally) must be assessed. Even more, the basic rationales of policy measures that are analysed must be compared to extract any competing views on cause-relationships, nature of innovation dynamics, problem perceptions etc. This is at the heart of the evaluation synthesis when considering a policy mix, as instruments tend to focus on one or two actor groups on the basis of heterogeneous institutional rationales of those institutions responsible for one specific measure.

Target groups and incentive schemes

Unless the evaluation synthesis focuses on the effects of a set of policies on one specific target group, it is indispensable to differentiate, as far as possible on the basis of existing evaluations, between the effects on different target groups. Furthermore, the various incentive schemes used for these target groups need to be characterised and their relation made obvious. Again, as with the measure targets, incentives can be complementary, re-enforcing, contradictory or doubling in the sense that in their co-existence measures produce windfall gains.

Integration of quantitative data

Evaluation synthesis is largely qualitative. Depending on the availability of quantitative evaluations, the evaluation synthesis can, however, also include quantitative data and translate and interpret it for the qualitative analysis. This means the evaluator needs to be capable of integrating and interpreting a numerical result of an evaluation, e.g. what the results of a scoring model for the performance criteria actually mean. For the target dimensions that are of interest for the evaluation synthesis one can formulate the benchmark and then assess (a) the importance of this target/dimension for the individual instrument and (b) the extent to which the goal was reached.

Context analysis

The characterisation of the policy measure also shall include, as shown above, the broader context of a measure. All evaluation syntheses need to make a context analysis or must include a secondary analysis of the context variables that influence the behaviour of actors. If the evaluations are sound on context analysis (see above), their findings can be included. However, the overall evaluation target may need an additional context analysis, both because of time (the evaluation synthesis takes place long after the context analysis for measures has been conducted) and because of the specific focus needed for the evaluation synthesis questions and scope.

Presentation of the findings

The presentation of the findings of an evaluation synthesis is not trivial. As policymakers may be confronted with assessments of their programme from a new perspective, they might question the study on many grounds. Thus it is indispensable that the evaluation is presented in a way that guarantees full transparency on the studies used and on the interpretations made on that basis. For an evaluation synthesis and the meta-evaluations they build upon the same standards applying to traditional evaluations, and the compliance with these standards must clearly be communicated. This also means that the selection and characterisation of policies based on the evaluations and the preceding quality check of the evaluations also must be clearly reported.

4.3.3 Complementary Activities – Inserting Interaction

As evaluation synthesis covers new ground in concentrating on interplays, gaps, redundancies and re-enforcing effects, the information stemming from existing evaluations may not be sufficient in some of the major research questions. For example, the co-ordination with complementary measures is more often than not neglected in evaluations of single measures. Furthermore, in some instances it may be necessary to

get a better in-depth understanding of specific measures or sets of measures, for example, those that have been rated as best practice in evaluations. Thus, the principle of interpreting and synthesising secondary data may be complemented by a set of additional activities such as case studies, interviews with strategic policymakers and programme managers and additional institutional analyses in order to understand coordination and interplay issues better.

A standard tool of evaluation synthesis should be, in any case, expert panels composed of key policymakers, ideally representatives of programmes evaluated and further strategic policy actors and stakeholders. These experts are not only best suited to assess and comment the results of an evaluation synthesis and to discuss findings and recommendations. Such a panel would also be a formative element in the evaluation synthesis, as it confronts policymakers with their contextual relevance and systemic position. Above all, not only the institution issuing an evaluation synthesis should be better informed, but also the individual policymakers at programme level.

4.3.4 Extension: Assessment of Evaluation Culture and Capabilities

In addition to the analyses of the content and quality of the evaluations and the nature and performance of policy measures, a further, simple analytical step will be to work towards a measure that describes the evaluation culture in specific systems.

There is no agreed definition for "evaluation culture" and different understandings and definitions are used in the literature (OECD 1998, Toulemonde 2000, GAO 2003, LL&A et al. 2006b). Combining and modifying the definition s of LL&A (2006a, p. 14) and OECD (1998, p. 5) in a "good evaluation culture" evaluations are regularly incorporated in the whole policy cycle (from design to implementation and follow-up activities), supported and taken into account by policymakers and administrators, demanded and taken into account by stakeholders, conducted by well trained, independent and credible experts using state-of-the-art and appropriate methods and adjusted to changing and divergent needs.

Our multi-module approach of assessing and using evaluations enables to contribute to the assessment of evaluation culture and capabilities in a given innovation system.[11] For any given innovation system, we can assess the relative importance of evaluation in the policy process. Following simple calculations can be done for the innovation sys-

11 A more traditional, but very enlightening case study approach on assessing evaluation culture was conducted by LL&A et al. 2006b, chapter 3.

tem and the types of policy under review: the share of programmes for which an evaluation is done, the share of programmes for which a sound evaluation exists, the share of evaluations that are published out of all evaluations conducted (as non-publication is an indication that something in the full process was sub-optimal).

However, the most important approach to assess evaluation culture would be the scoring model of quality that was introduced above. This scoring model and the weighing of criteria allow a differentiated and tailored approach to the assessment of quality, utility and uptake of evaluations. For an assessment of evaluation culture, however, one would have to extend the list of criteria used in exhibit 2 and include a more comprehensive list. In addition to the full list of standards of the American Evaluation Society, an appropriate list of criteria is also provided by the Austrian Evaluation Platform (efteval 2003 and by Stufflebeam 2000, p. 113 ff). As with the evaluation synthesis on policy measures above, an interaction with policymakers and evaluators to check and comment this assessment would be an important additional step.

Evaluation culture and capabilities is important information for the assessment of policymaking in innovation systems, thus it is a complementary aspect of evaluation synthesis, informing not only about the "what" and the effect, but also about the "how" of innovation policy. A second value of this assessment is that it can also be fed back into the meta-analysis in a very basic way. Evaluation culture in a given system can in itself become a variable in our meta-analysis and we subsequently can analyse the impact of evaluation culture and capabilities on the effects of policy measures. The assumption to test would be that the better the evaluation culture, the more appropriate and relevant the policy measures and the better their effects.

4.4 Benefits and Limits

4.4.1 Benefits

The evaluation synthesis gains new systems insights on the basis of existing knowledge by systematic and cost-efficient desk work. The most important benefit of an evaluation synthesis is to shed light on systems issues, the relative position and contribution of individual programmes and the appropriateness of a policy mix. Furthermore, the evaluation synthesis opens up the policy discourse from the programme to the systems level and has as such a strong formative element.

By systematically taking into account the various evaluations on certain issues, one can aggregate and synthesise overlapping questions and context descriptions discussed by different evaluators and analysed by means of different methodological ap-

proaches. Thus there is more robustness in findings on variables such as innovation dynamics, the significance of certain policy measures and context variables.

Furthermore, implicitly or, if included in the evaluation synthesis objective, explicitly individual policies are benchmarked in an evaluation synthesis. A well contextualised evaluation synthesis can help to overcome the limits of traditional benchmarking that are often criticised for not taking the systems perspective into account (Paasi 2005; Lundvall / Tomlinson 2001). The added value of evaluation synthesis on the programme level is that it is much more contextual than a meta-analysis and thus, for a given situation in a given system, can render very concrete insights and recommendations. Policy measures might be improved simply by better understanding how they interrelate with others. Thus, the assessment of individual programmes might be modified in the light of the evaluation synthesis results. This would cure an ill of most evaluations of individual programmes which often neglect the systemic position of the measure.

In the unlikely case that in RTDI policy enough studies exist for one single measure to render an evaluation synthesis possible and sensible, such an analysis would put assessments and recommendations on a higher level.

Finally, the evaluation synthesis informs us about evaluation culture and capabilities and also reveals gaps in the knowledge of policy performance and implementation. This adds another important formative element beyond individual measures.

4.4.2 Limits

The most important limit of evaluation synthesis concerns the pre-conditions to realise its benefits: we can only synthesise what we have. But there might be a systematic or random bias towards a certain set of instruments, we might only get to know the evaluations that have been positive, rather than having access to all evaluations etc. The worse the evaluation culture as defined above, the less likely that the results of an evaluation synthesis will render robust and valid results. This is why meta-evaluation is so important and why we believe that a sound evaluation synthesis may be less costly than a large systems review. Only if we are very transparent and conscious about the quality of the data basis for an evaluation synthesis can we claim to obtain valuable results. If evaluations are lop-sided towards the instruments evaluated well, towards those that are easier to evaluate or towards those of one single, evaluation-friendly ministry etc., the results of the evaluation synthesis are lopsided, too. This has to be taken into account in the overall synthesis and the recommendations. It is mandatory that the reporting of the results must be as transparent as possible. One further remedy

for this problem could be some form of overall expert judgement based on multi-perspective interviews.

The evaluation synthesis can only answer questions that have somehow also been addressed in the individual evaluations. If a question is relevant at the systems level that was not relevant at the programme level (and thus not reflected in the evaluation of the programme(s)), an evaluation synthesis cannot provide deeper insights without additional analysis going beyond the reliance on existing data and interpretation, such as an expert panel or some expert interviews. One such issue that is important for evaluation synthesis, but is mostly not dealt with in evaluations is that interfaces and the interplay between measures is very often not analysed in individual evaluation studies. While the lack of this perspective is in itself an important result of a meta-analysis and an evaluation synthesis, it still limits the analyses in an evaluation synthesis.

One final limit for evaluation synthesis, in fact a reason why those analyses are not conducted very often, may lie in institutional opposition to such analyses, as in most countries we find horizontal and – in federal systems – vertical fragmentation of institutions that are responsible for RTDI policy. The institutional incentive to have a full systems portfolio evaluated are often lacking, and a systems perspective on individual measures may, in effect, limit the degree of freedom for the policymakers responsible for those measures, as individual and system agendas might not be complementary.

5 Conclusion – Strategic Benefits of Secondary Analysis

This paper developed and discussed two secondary analyses that both build upon existing evaluations in order to exploit the knowledge gained in the large number of existing evaluations for deeper and broader insight. Two types of secondary analysis have been discussed

(1) the quantitative method of meta-analysis on the level of measures to shed light on the functionalities of types of policy measures and on the intermediary variables influencing policy effects and

(2) the qualitative evaluation synthesis that aggregates the findings of individual measures in order to assess interplay and systemic performance. The benefit of the evaluation synthesis is that it helps understand the relative role of policies in a given system context and in the interplay with other approaches.

The basis to conduct both of these variants are meta-evaluations that serve to characterise, check and select existing evaluations, and which in itself can achieve insights into evaluation cultures.

The strategic benefits of both approaches are clear. Through the meta-evaluation the system is informed about its evaluation culture and capabilities. Through the meta-analysis, especially policymakers can gain – at rather low cost – in-depth knowledge of the conditions and potentials of various types of policy measures, checked for different context variables. Through evaluation synthesis policymakers and other stakeholders gain insights into the interplay of policy mixes and policymakers can better position themselves in the web of measures in a given innovation system.

It is a shortcoming of RTDI policymaking that the potentials that lie in these approaches are not better explored. As the GOA pointed out already in 1992, "no single study, no matter how good, can have this kind of power" (GOA 1992, p. 6). We are well aware of all the bottlenecks for such broad analyses, beginning with the level of an evaluation culture that is needed that produces a sufficient number of sound evaluations to be taken up for the analyses. However, strategic intelligence and policymaking alike cannot only rely on ad hoc evaluations assessing isolated programmes, and foregoing an important source of systemic formative evaluation. Rather than sticking to this practice, innovation systems should be better equipped with the preconditions for secondary analysis. Better evaluation cultures are called for.

This article does not claim to have produced a manual, rather it is a conceptual skeleton that lays out principles to be used if one intends to go in the direction outlined. Areas of application are many. To conclude, two promising fields may be suggested. For example, for more than 20 years now each OECD country has programmes to foster co-operation between universities and companies. There are evaluations enough to select a sufficient number of cases and conduct a meta-analysis in order to better understand certain context and instrument variables. For evaluation synthesis – as one example among many potential fields of application – one could imagine that the shaping of technological innovation systems across Europe could be accompanied by an evaluation that asks for the overall effect of policies in European countries for these technologies and thus to see what a supranational approach could add to the system to close gaps and remove bottlenecks.

Literature

Arnold, E. (2002): Evaluation research and innovation policy. A systems world needs systems evaluations. In: Research Evaluation 13 (1), 3-17.

Arnold, E.; Clark, J.; Muscio, A. (2005): What the evaluation record tells us about European Union Framework Programme performance. In: Vonortas, N.S.; Hinze, S. (ed.), Special issue of Science and Public Policy on "Evaluation of European Union Framework Programme: The 2004 Five Year Assessment"; 32 (5), October 2005, 385-397.

Arnold, E.; Kuhlmann, S; van der Meulen, B. (2001): A Singular Council. Evaluation of the Research Council of Norway, Brighton.

Beelmann, A.; Bliesener, T. (1994): Aktuelle Probleme und Strategien der Metaanalyse. Psychologische Rundschau 45, 211-233.

Beywl & Associates GmbH (2004): Glossar wirkungsorientierte Evaluation. Univation-Institut für Evaluation, Cologne.

Blanpied, W. A. (2004): Achievements of the Science and Technology Basic Plans. Impressions of a Sympathetic Foreigner. Presentation at the NISTEP international Workshop on Comprehensive Review of Japan's Science and Technology Basic Plans. Tokyo, Japan, September 13-14, 2004.

Borgmann, M. (2005): Evaluation Synthesis zu Angeboten der Wissenschaftskommunikation im Rahmen der Evaluation des "Jahrs der Technik 2004". Studie im Auftrag des BMBFM, Cologne.

Borrás, S. (2004): System of Innovation Theory and the European Union. In: Borrás, S. (ed.): Special Issue of Science and Public Policy on a European system of innovation 31 (6), 425-433.

Borrás, S. (ed.) (2004): Special Issue of Science and Public Policy on a European system of innovation 31 (6).

Carlsson, B. (ed.) (1995): Technological Systems and Economic Performance: The Case of Factory Automation. Kluwer Academic Publishers, Boston.

Carlsson, B. (ed.) (1997): Technological Systems and Industrial Dynamics. Kluwer Academic Publishers, Boston.

Carlsson, B. et al. (2002): Analytical approach and Methodology. In: Carlsson, B. (ed.), Technological Systems in the Bio Industries: An International Study. Kluwer Academic Publishers, Boston.

Carlsson, B.; Jacobsson, S.; Magnus Holmén, M.; Rickne, A. (2002): Innovation systems: analytical and methodological issues. In: Research Policy 31, 233–245.

Carlsson, B.; Stankiewicz, R. (1991): On the nature, function, and composition of technological systems. Journal of Evolutionary Economics 1 (2), 93–118.

Cooke, P. et al. (eds) (1998): Regional Innovation Systems: The Role of Governances in a Globalized World, Routledge.

Cooksy, L. J.; Caracelli, V. J. (2005): Quality, Context and Use. Issues in Achieving the Goals of Meta-Evaluations. In: American Journal of Evaluation, 261, March 2005; 31-42.

DeCoster, J. (2004): Meta-analysis Notes. Retrieved Oct 3, from http://www.stathelp.com/notes.html

Drinkmann, A. (1990): Methodenkritische Untersuchungen zur Metaanalyse. Weinheim: Deutscher Studien Verlag.

Edler, J.; Amanatidou, E.; Bührer, S; Cunningham, P; Polt, W; Guy, K; von Oertzen, J. (2006a): Perspectives On Evaluation and Monitoring. Project Proposal of Fraunhofer ISI, PREST, Joanneum, Wise Guys and Atlantis for the European Commission DG Enterpise. Project awarded November 2006.

Edler, J.; Dreher, C.; Ebersberger, B. (2006): Market and system failures and policy instruments in the technology cycle. In: Jochem, E. et al. (2006): Developing an assessment framework to improve the efficiency of R&D and the market diffusion of energy technologies (EduaR&D). Project Report to the BMWi (forthcoming).

Edquist, C. (ed.) (1997): Systems of Innovation: Technologies, Institutions, and Organizations. Pinter, London.

Eisend, M. (2004): Metaanalyse - Einführung und kritische Diskussion. Diskussionsbeiträge des Fachbereichs Wirtschaftswissenschaften der FUB, 2004/8, Berlin.

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

FFG; Fraunhofer ISI, MPFL, WWTF (2004): RoadMAP. Good practices for the management of Multi Actors and Multi Measures Programmes (MAPs) in RTDI policy.

Fleiss J.L. (1993): The statistical basis of meta-analysis. Statistical Methods in Medical Research 2, 121–145.

FTEVAL (Plattform Forschung- und Technologieevaluierung) (2003): Evaluation Standards in Research and Technology Policy; Vienna (http://www.fteval.at/standards /Gesamt% 20Standards%20E.pdf)

GAO (United States General Accounting Office) (1992): The evaluation synthesis. Revised March 1992. GAO/PEMD-10.1.6, Washington D.C.

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

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

Georghiou L.; Roessner, D. (2000): Evaluating technology program: tools and methods. Research Policy (29) 4-5, 6.

Georghiou, L.; Amanatidou, E.; Belitz, H.; Cruz, L.; Edler, J.; Edquist, C.; Granstrand, O.; Guinet, J.; Leprince, E.; Orsenigo, L.; Rigby, J.; Romanainen, J.; Stampfer, M.; van den Biesen, J. (2003): Improving the Effectiveness of Direct Public Support Measures to Stimulate Private Investment in Research. Report of the ETAN Working Group on Direct Measures for Directorate General Research. European Commission, Brussels.

Glass, G.V. (1976): Primary, Secondary, and Meta-Analysis of Research. Educational Researcher 5, 3-8.

Hedges, L. V.; Gurevitch, J.; Curtis, P. S. (1999): The Meta-Analysis of Response Ratios in Experimental Ecology. Ecology 80, 1150-1156.

Hekkert. M.P.; Suurs, R.A.A.; Kuhlmann, S.; Smits, R.E.H.M., (2006): Functions of Innovation Systems: A new approach for analysing technological change. In: Technological Forecasting and Social Change (forthcoming).

Jochem et al. (2006) (ed.): Developing an assessment framework to improve the efficiency of R&D and the market diffusion of energy technologies (EduaR&D). Project Report to the BMWi (forthcoming).

Kondo, M (2004): Highlights of the Comprehensive Review of Japan's Science and Technology Basic Plans. Presentation at the NISTEP international Workshop on Comprehensive Review of Japan's Science and Technology Basic Plans. Tokyo, Japan, September 13-14, 2004.

Koschatzky, K.; Lo, V. (2005): Innovationspolitik in den neuen Ländern. Bestandsaufnahme und Gestaltungsmöglichkeiten. Stuttgart: Fraunhofer IRB Verlag.

Kuhlmann, S. (2001): Management of Innovation System. The Role of Distributed Intelligence. Nijmwegen Lectures on Innovation Management, Antwerp/Apeldorn.

Kuhlmann, S. et al. (1999): Improving Distributed Intelligence in Complex Innovation Systems. Final report to the ASTPP. Karlsruhe, June 1999.

Light, R. (1984): Six evaluation issues that synthesis can resolve better than single studies. In: Yeaton, W.H.; Wortmann, P.M. (eds.): Issues in data synthesis. New Directions for Program evaluation, 24, San Francisco, 57-73.

LL&A, PREST, ANRT, Reidev Ltd (2006a): SMART INNOVATION: A practical Guide to evaluating innovation programmes. January 2006.

LL&A, PREST, ANRT, Reidev Ltd (2006a): Supporting the monitoring and evaluation of innovation programmes. Final report, January 2006.

Lundvall, B.-A. (ed) (1993): National Innovation Systems: Towards a Theory of Innovation and Interactive Learning. Pinter, London.

Lundvall, B.-A.; Tomlinson, M. (2001): International benchmarking as a policy learning tool. In: Lundvall, B.-A. et al. (ed.): The New Knowledge Economy in Europe. Cheltenham; 203-231.

Malerba, F. (2002): Sectoral systems of innovation and production. Research Policy 31, 247-264.

Malerba, F. (2004):, Sectoral Systems of Innovation: Basic Concepts. In: Malerba, F. (ed.), Sectoral Systems of Innovation. Concepts, Issues and Analyses of Six Major Sectors in Europe. Cambridge, 9-41.

Meyer-Krahmer, F.; Kuntze, U. (1992): Bestandsaufnahme der Forschungs- und Technologiepolitik. In: Grimmer et al. 1992, 95-118.

Miller, N.; Pollock, V.E. (1994): Meta-analytic synthesis for theory development. In: Cooper, H.; Hedges, L.V. (eds.), Handbook of research synthesis. New York: Russel Sage Foundation, 457-483.

Nelson, R. R. (ed.) (1993): National Innovation Systems. A Comparative Analysis, Oxford University Press, New York/Oxford.

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

Paasi, M. (2005): Collective benchmarking of policies: an instrument for policy learning in adaptive research and innovation policy. In: Science and Public Policy 32 (1); 17-27.

Ruegg, R (2005): The advanced Technology Program's evaluation plan & progress http://www.atp.nist.gov/eao/7th-iftm.htm, last update April 2005. Access November 2006.

Ruegg, R.; Feller, I. (2003): A Toolkit for Evaluating Public R&D Investment Models, Methods, and Findings from ATP's First Decade. Report to the Department of Commerce, Washington D. C. July 2003. (http://www-15.nist.gov/eao/gcr03-857/contents.htm)

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

Scriven, Michael (1991): Evaluation Thesaurus, 4th edition, Newbury Park.

Stanley, T. D. (2001): Wheat from Chaff: Meta-Analysis as Quantitative Literature Review. Journal of Economic Perspectives 15, 131-150.

Stufflebeam, D.L. (2000): The Methodology of Meta-evaluation as Reflected in Meta-evaluations by the Western University Evaluation Center. Journal of Personnel Evaluation in Education 14 (1), 95-125.

Stufflebeam, D.L. (2001): The Meta-evaluation Imperative. American Journal of Evaluation 22 (2), 183-209.

Toulemonde, Jacques (2000): "Evaluation Culture(s) in Europe : Differences and Convergence between National Practices." Vierteljahrshefte zur Wirtschaftsforschung/Quarterly Journal of Economic Research. DIW Berlin, German Institute for Economic Research, Vol. 69(3), 350-357.

Vonortas, N.S.; Hinze, S. (ed.) (2005): Special issue of Science and Public Policy on "Evaluation of European Union Framework Programme: The 2004 Five Year Assessment", 32 (5), October 2005.

Widmer, T. (1996): Meta-Evaluation. Kriterien zur Bewertung von Evaluationen. Bern u.a.

Widmer, T; Beyl, W. (2000): Die Übertragbarkeit der Evaluationsstandards auf unterschiedliche Evaluationsfelder. In: Joint Committee on Standards for Educational Evaluation (ed.). Handbuch der Evaluationsstandards; Opladen; 234-257.

Appendix: Evaluation Standards

The Standards for Evaluation of the German Evaluation Society (DeGEval) as one guideline for the quality check and assessment of evaluations to be used in the Secondary Analysis

Source: http://www.degeval.de/calimero/tools/proxy.php?id=72

UTILITY
Utility standards are intended to ensure that an evaluation is guided by both the clarified purposes of the evaluation and the information needs of its intended users.

U1 Stakeholder Identification
Persons or groups involved in or affected by the evaluand should be identified, so that their interests can be clarified and taken into consideration when designing the evaluation.

U2 Clarification of the Purposes of the Evaluation
The purposes of the evaluation should be stated clearly, so that the stakeholders can provide relevant comments on these purposes, and so that the evaluation team knows exactly what it is expected to do.

U3 Evaluator Credibility and Competence
The persons conducting an evaluation should be trustworthy as well as methodologically and professionally competent, so that the evaluation findings achieve maximum credibility and acceptance.

U4 Information Scope and Selection
The scope and selection of the collected information should make it possible to answer relevant questions about the evaluand and, at the same time, consider the information needs of the client and other stakeholders.

U5 Transparency of Values
The perspectives and assumptions of the stakeholders that serve as a basis for the evaluation and the interpretation of the evaluation findings should be described in a way that clarifies their underlying values.

U6 Report Comprehensiveness and Clarity
Evaluation reports should provide all relevant information and be easily comprehensible.

U7 Evaluation Timeliness
The evaluation should be initiated and completed in a timely fashion, so that its findings can inform pending decision and improvement processes.

U8 Evaluation Utilisation and Use
The evaluation should be planned, conducted, and reported in ways that encourage attentive follow-through by stakeholders and utilisation of the evaluation findings.

FEASIBILITY
The Feasibility Standards are intended to ensure that an evaluation is planned and conducted in a realistic, thoughtful, diplomatic, and cost-effective manner.

F1 Appropriate Procedures
Evaluation procedures, including information collection procedures, should be chosen so that the burden placed on the evaluand or the stakeholders is appropriate in comparison to the expected benefits of the evaluation.

F2 Diplomatic Conduct
The evaluation should be planned and conducted so that it achieves maximal acceptance by the different stakeholders with regard to evaluation process and findings.

F3 Evaluation Efficiency
The relation between cost and benefit of the evaluation should be appropriate.

PROPRIETY
The propriety standards are intended to ensure that in the course of the evaluation all stakeholders are treated with respect and fairness.

P1 Formal Agreement
Obligations of the formal parties to an evaluation (what is to be done, how, by whom, when) should be agreed to in writing, so that these parties are obligated to adhere to all conditions of the agreement or to renegotiate it.

P2 Protection of Individual Rights
The evaluation should be designed and conducted in a way that protects the welfare, dignity, and rights of all stakeholders.

P3 Complete and Fair Investigation
The evaluation should undertake a complete and fair examination and description of strengths and weaknesses of the evaluand, so that strengths can be built upon and problem areas addressed.

P4 Unbiased Conduct and Reporting
The evaluation should take into account the different views of the stakeholders concerning the evaluand and the evaluation findings. Similar to the entire evaluation process, the evaluation report should evidence the impartial position of the evaluation team. Value judgments should be made as unemotionally as possible.

P5 Disclosure of Findings
To the extent possible, all stakeholders should have access to the evaluation findings.

ACCURACY
The accuracy standards are intended to ensure that an evaluation produces and discloses valid and useful information and findings pertaining to the evaluation questions.

A1 Description of the Evaluand
The evaluand should be described and documented clearly and accurately, so that it can be unequivocally identified.

A2 Context Analysis
The context of the evaluand should be examined and analysed in enough detail.

A3 Described Purposes and Procedures
Object, purposes, questions, and procedures of an evaluation, including the applied methods, should be accurately documented and described, so that they can be identified and assessed.

A4 Disclosure of Information Sources
The information sources used in the course of the evaluation should be documented in appropriate detail, so that the reliability and adequacy of the information can be assessed.

A5 Valid and Reliable Information
The data collection procedures should be chosen or developed and then applied in a way that ensures the reliability and validity of the data with regard to answering the evaluation questions.

A6 Systematic Data Review
The data collected, analysed, and presented in the course of the evaluation should be systematically examined for possible errors.

A7 Analysis of Qualitative and Quantitative Information
Qualitative and quantitative information should be analysed in an appropriate, systematic way, so that the evaluation questions can be effectively answered.

A8 Justified Conclusions
The conclusions reached in the evaluation should be explicitly justified, so that the audiences can assess them.

A9 Meta-evaluation
The evaluation should be documented and archived appropriately, so that a meta-evaluation can be undertaken.