technopolis |group|

24 March 2014

In search for a benchmark of impact, effectiveness and efficiency of innovation instruments

Management Summary

A report for the TAFTIE Task Force on Benchmarking Impact, Effectiveness and Efficiency (TFBIEE)

technopolis |group|

In search for a benchmark of impact, effectiveness and efficiency of innovation instruments

Management Summary

technopolis |group| March 2014

Patries Boekholt (project leader)
Matthias Ploeg
Erik Arnold
Flora Giarracca

with contributions from:
Peter Kolarz, Bea Mahieu, Jacek Walendowski and Katrin Männik

Management Summary

This report for the TAFTIE Task Force on Benchmarking Impact, Effectiveness and Efficiency (TFBIEE) looks at the accomplishments of innovation policy instruments in an international context. The Task Force was set up to gain a combined insight in innovation agencies' impact, effectiveness and efficiency. In order to fulfil TAFTIE's ambition to demonstrate the added value of innovation agencies to society, it is important to understand and demonstrate how the policies these agencies implement have an impact. The main purpose of this project is to learn from each other and improve the agencies' impact. A comparison or benchmark between the agencies in the Task Force would contribute to joint policy learning.

As agencies have quite different mandates, tasks, resources and policy portfolios, a comparison on agency level will be distorted by these significant differences. In addition, it is mostly at programme level that impacts are established through evaluation studies. This study was a first experiment to assess to what extent it is possible to compare effectiveness between specific types of innovation instruments across different countries. At the outset it was clear that the design of the instruments and the dissimilar contexts in which agencies and instruments operate would make this an almost impossible task. In an ideal case we could compare the claims made on impacts as reported in evaluation studies.

In order to maximise comparability in this benchmark, TFBIEE decided to select a number of specific innovation instruments that have similar goals and design features.

The four categories of instruments chosen were:

1. Grant schemes for R&D with business as beneficiaries
2. Grant schemes for R&D for collaborative research with business and research institutions as beneficiaries
3. Integral cluster and competence centre schemes
4. Innovation Voucher schemes for business

By choosing instruments with similar rationales to compare with each other, the design of the programme would be less of a distorting factor in the cross-border comparison.

This assignment is based on three research questions:

- Is it possible to benchmark impact/effectiveness of policy instruments in an internationally comparable way?
- How do we estimate the impact/effectiveness of these instruments? Are agencies in line with (inter)national handbooks and guidelines in this matter?
- Is it possible to benchmark innovation agencies on key figures of the implementation process?

In order to answer the first research question - whether it is possible to compare effectiveness of similar instruments - a necessary preceding step was to establish whether an objective and robust assessment of the impacts was made for each of these instruments. This is commonly done through evaluation studies.

Thus the first step of our assignment – dealing with the second research question – was to review whether the impact assessments were conducted according to international 'good practice'. We did this by synthesising the existing (generic) guidelines for good evaluations, literature on evaluation and combined this with other

evaluation experience from the innovation field. This was translated into a practical tool: the Evaluation Reference Model. The Evaluation Reference Model, which was developed for this project based on existing guidelines and handbooks, can be seen as an elaborated checklist for the key steps to take before, during and after an evaluation.

It was however not intended to be another handbook on evaluation, but a reference guide to use alongside the many existing guidelines and handbooks. The reference model seems to have worked well to make an assessment of the strengths and weaknesses of the evaluation studies in the benchmark, even though these assessments need to take account of many contextual factors.

A considerable part of the reference model is dedicated to the choice of evaluation methodology. However, methodology is an important but not sufficient criterion to assess the quality of evaluations. State-of-the-art evaluations are based on wider considerations that impact the methodology design and implementation of evaluation tools such as utility of the evaluation for intended users, involvement of stakeholders in the evaluation, clarity and transparency in the analysis. This is why the Evaluation Reference Model presents a **process approach to evaluation** and is structured around six key steps:

- **Step 1**: Definition of the Programme Logic Model
- **Step 2**: Definition of the evaluation objectives and questions
- **Step 3**: Preparation of the evaluation
- **Step 4**: Identification of appropriate methodology for analysis (state-of-the-art methodology mix by type of instrument)
- **Step 5**: Conclusions and reporting of the evaluation
- **Step 6**: Apply the lessons from the evaluation

For each of these steps a list of key assessment criteria are identified that constitute the basis for best evaluation practices.

The subsequent step was to compare the existing evaluation studies of the benchmark instruments to the reference model. This led to our **first benchmark** of the **evaluation approach** used for the set of four types of instruments. Each step in the reference model has a number of criteria against which evaluation practices have been scored. A qualitative assessment was given of the alignment of the evaluation studies with the reference model, taking into account specific context variables, such as the purpose of the study and whether it was a mid-term or an ex-post evaluation. For example for mid-term reviews the expectations regarding to demonstrating impacts are different than for an ex-post impact assessment.

With the information of the evaluations at hand a second step was to log and compare the effects and impacts that were found for each of these instruments –as reported in the evaluation studies. We compared instruments within their category across the ten countries. In order to do this an overview was made of the main indicators used to report on effectiveness. To understand the influence of the design of the instruments we also looked at key design features such as the size of the programme, the funding mechanisms and the target groups. With this overview of instrument designs and reported effects, the task was to conduct the **second benchmark** of the **effectiveness of instruments** within their category. We firstly tried to identify whether similar indicators were used for measuring success of the programme. We subsequently looked for any quantifiable statements on impacts that could be compared across instruments. And as a last task we looked for any statements of economic impacts beyond the direct participants.

A **third benchmark** focused on the **administrative efficiency** of the instruments within their category. For each instrument we collected information on an agreed set of indicators for programme implementation. This included handling costs, success rates, time-to-contract, complaints and for voucher schemes return rates.

Thus, the **first benchmark** consisted of comparing the evaluations and studies made of the 28 instruments across ten countries. We came to the conclusion that comparing how we evaluate our programmes has been informative and will be useful for future practices. It helps to start a discussion amongst agencies on the best ways to establish effectiveness and impacts and help the agencies internally to rethink their evaluation approaches. The reference model and assessment grids can be used to expand the number of instruments and evaluations to be benchmarked in the future. The assessment was made on the basis of a quick scan of the programmes based on existing material so the intention was not to identify the best performing instrument or agency.

The review of evaluations and their programme documentation shows that a major bottleneck for evaluation starts with the **broadly defined general programme objectives** that have not been operationalised into more specific objectives, on the basis of which one can decide whether a programme has been successful or not. This poses the evaluators with the problem of not being able to give clear answers to whether the programme has achieved its goals. We have seen only a few examples of evaluations, which have tried to reconstruct the objectives and the logic intervention model of the programme. We have not encountered any evaluation that is focusing on a set of key performance indicators that have been defined by the programme managers.

In the **definition of evaluation objectives** we observed many cases where the scope of evaluation questions was limited, mostly to the question on demonstrating the impacts. While it is legitimate to narrow the scope of a study down to one or two evaluation questions as long as this is done deliberately and as long as programme owners realise they will not get a full picture of whether a programme's objectives are still relevant, whether the programme contributes to the problem it addresses and more importantly why it is not fulfilling its full potential if the impacts are not impressive.

We have seen a - small - number of good examples where evaluations also make a **portfolio analysis** and position the instrument that is analysed within a broader package of policies and the innovation system.

In the preparation of the evaluation the **availability of data** is clearly an issue for many evaluations, although rarely do the reports make a critical analysis of the existing datasets. An increasing number of countries have developed a good set of (micro-level) baseline data on the companies participating and not participating in the programmes. This is not common practice in all countries as we have understood. A key task of the TAFTIE agencies in countries where this is not yet allowed would be to arrange the use of a wider set of data on behalf of the evaluations.

A key element for good evaluations is the choice of the methodology mix. The assessment of the evaluation studies show that there is a lot of good practice across the TAFTIE countries, but at the same time there is still a lot of room for improvement at the level of individual evaluations. The methodology mix should be fit for purpose, dependent on the timing of the evaluation and the evaluation questions asked. There is not one recipe for a good methodology mix. In the benchmark we have seen the largest issues with a **lack of triangulation**, where conclusions were drawn using one methodology only or making general statements from a non-representative sampling of beneficiaries.

Using surveys with control groups is also becomes more common and particularly in the business grant and voucher schemes a majority of studies use these methods. A number of agencies are experimenting with better ways to do counterfactual analysis using experimental econometrics methods, although this is still a minority in the set of evaluation studies. However, the examples also show that to get the full picture, these econometric analyses need to be part of a wider methodology mix to obtain an understanding why certain effects occur or not occur.

Regarding evaluation reporting and conclusions the issues found were mostly about the clarity of the report and about linking the evidence from the study to the final conclusions. In well written evaluation reports the evidence from the main chapters was summarised and presented in such a manner that the arguments for the conclusions were clearly derived on those findings.

In summary the key evaluation issues that would need to be addressed by the agencies are:

- Devote more efforts to defining the programme objectives and goals in a sufficiently concrete manner in order to render the evaluation of achievements and success more in line with the programme logic model. Develop an evaluation framework already at the start of the programme
- Define the data requirements for an evaluation at the start of the programme and develop baseline and monitoring data that would be needed for later evaluations
- Adapt the set of evaluation questions to the timing of the evaluation and be realistic about the time line at which one can expect to measure impacts at company level and as spill-overs to the rest of the economy and society
- Work on the access to appropriate micro-level data that can be linked to the instruments' population as well as any statistically constructed control group. If confidentiality is an issue, collaborations with national statistics offices have been successfully developed by a number of the TAFTIE agencies
- Insist from your evaluation team a methodology mix that is fit for purpose and allows for sufficient triangulation between the methods used.
- Insist also on where possible in view of the data availability and the instrument's target group on some form of counterfactual analysis for the assessment of net effects
- Insist from the evaluation that the conclusions can be easily derived from the evidence gathered and that all the methods used are transparent

The s**econd benchmark**, comparing the impacts between instruments across the agencies borders appeared to be more problematic. On the research question '*is it possible to benchmark impact/effectiveness of policy instruments in an internationally comparable way?*' we would simply have to answer not with the present differences in evaluation approaches and methods used. The most important reason is that a majority of the evaluations focus on inputs, throughputs and outcomes of programmes, while having quite weak evidence on the impact on innovation capabilities, increase in competiveness or economic benefits of the programmes. The second reason is that the evidence that is presented is based on different sets of empirical data, indicators for success and data gathering methods. But we have also identified some opportunities and possibilities for a better comparison in the future. At best we can compare outcomes of similar questions in surveys across similar programmes, even though survey questions are also opinion based and have the usual bias issues. In the grant and collaborative R&D schemes this could be achieved if the agencies would agree to align some of their survey evaluation questions to the participants. This would then still need additional evaluation methods to interpret and explain the results in the local context. Having a more standardised way of defining control groups to address the additionality issue and include these control groups in the surveys would already be an important next step to take across all countries. In those cases – for the business grants and R&D collaborative grant schemes - where external data are used to compare the participants with a control group, an agreement could be made to use a set of common indicators. CIS data can be used as a common dataset across countries.

As we have not been able to take the step to compare similar outcomes of evaluation studies, the question whether it is methodologically possible to control for differences

in the context in which these instruments operate, has not yet been tested. Given the complexity of the integrated cluster and competence centre programmes, in our view it only makes sense to compare these type of programmes across countries if these programmes have had a long history of at least 10 years and if good baseline studies exist to understand the evolution of the clusters in time. For these instruments rather than comparing evaluations, it seems more fruitful to exchange expert knowledge on how these centres or clusters can be best organised and managed, how to get the best interaction between public and private actors and what formal arrangements should support these initiatives. A cross-border comparison of effectiveness could perhaps be better done at centre and cluster level (matching similar initiatives) rather than at programme level, as the set of clusters and centres within a programme will be too disparate.

On the **third research question** (*is it possible to benchmark innovation agencies on key figures of the implementation process?)* the short answer is yes we have managed to compare on a number of key indicators for implementation. The immediate caveat needs to be made that understanding what these figures really mean and tell us is not straightforward. For instance comparing handling costs across agencies using different accountancy systems and different systems to allocate costs to programmes needs to be taken with great caution. And in addition it does not give an answer to the question how the handling costs are related to the overall programme efficiency and effectiveness. Figures on time-to-contract for instance are useful to start a learning process amongst agencies to see how an optimal process could be achieved. Voucher return rates also give an insight in how well the instrument is received by the target group.

In summary, while the project was aware of the many challenges and bottlenecks of a cross agency benchmark it has provided a basis for further learning, particularly on improving the evaluation practices and achieving more robust findings on the achievements of innovation instruments. It has provided a good basis for next steps that TAFTIE can take for policy learning to improve innovation instruments and as a result increase their societal impact.

technopolis |group| The Netherlands
Herengracht 141
1015 BH Amsterdam
The Netherlands
T +31 20 535 2244
F +31 20 428 9656
E info.nl@technopolis-group.com
www.technopolis-group.com