EUROPEAN COMMISSION

DG Competition

Evaluation in the field of State aid

Draft METHODOLOGICAL GUIDANCE PAPER

- concepts and recommendations -

November 2013

Disclaimer: This document does not represent an official position of the European Commission. It is a tool to explore the views of interested parties. The suggestions contained in this document do not prejudge the form or content of any future proposal by the European Commission.

Contents

1 Introduction

Member States provide State aid for a wide variety of policy objectives, for instance, to reduce regional disparities within the country, to promote R&D and innovation activities, or to promote the adoption of a high level of environmental standards. The purpose of EU State aid control is to allow these types of State intervention, while making sure that the distortions of competition and trade in the EU remain limited.

In determining which types of aid are compatible with the common market, the EU State aid rules tend to be based on a system of *ex-ante* scrutiny: schemes are approved on the basis of pre-defined assessment criteria on the assumption that, as long as the schemes comply with these assessment criteria, their overall balance will be positive. Typically, this assessment of schemes is performed without the possibility to evaluate their actual impact on the markets over time[1].

Ex-post evidence on what has been really achieved with public funds or on the actual impact of State aid on competition has thus so far received relatively limited emphasis in the implementation of EU State aid rules. Measured impacts and lessons learnt are however essential for decision makers, both at the Member State and EU level. They allow ensuring more effective and less distortive measures, justifying public spending to the taxpayers as well as improving the efficiency of public interventions.

Already at present, a number of countries evaluate their subsidy measures, even if not always on a regular basis[2]. Some Member States also launched "spending reviews" and are considering the opportunity to review their aid policies so as to maximise their effectiveness in a context of fiscal consolidation[3]. In a similar vein, EU spending (*inter alia* EU structural and cohesion funds, employment aid) is subject to systematic *ex-ante* and *ex-post* evaluation to estimate the leverage effect, i.e. how much private spending was generated with the help of public funds in compliance with the applicable Regulation[4] and in line with the Guidance documents published by the Commission services[5].

The purpose of this methodological guidance paper is to help public authorities conduct ex-post evaluation of State aid schemes. This document advocates for best practices in order to

---

1 At the same time, aid schemes account for the vast majority of all granted aid: according to the 2012 Scoreboard data, approved aid schemes represent 23% of all aid measures and 55% of aid amounts, and a further set of block exempted schemes represent 63% of all aid measures and around 32% of aid amounts.

2 For instance, in several Member States, state aid reports are regularly prepared for the Court of Auditors or the Parliament.

3 Cf. for instance the Report commissioned to Professor F. Giavazzi by the Italian government, or the Report commissioned J-J. Queyranne, J-P. Demaël and P. Jurgensen by the French government.

4 Proposal for a Regulation of the European Parliament and of the Council laying down common provisions on the European Regional Development Fund, the European Social Fund, the Cohesion Fund, the European Agricultural Fund for Rural Development and the European Maritime and Fisheries Fund covered by the Common Strategic Framework and laying down general provisions on the European Regional Development Fund, the European Social Fund and the Cohesion Fund and repealing Council Regulation (EC) No 1083/2006. COM (2011) 0615 f/2.

5 Guidance documents on monitoring and evaluation for ERDF, Cohesion Funds or European Social Fund.

foster a common approach among Member States, thereby ensuring high quality and comparable evaluations between Member States.

The merits of conducting evaluations will materialise within a few years, when the first reports are ready and their findings and recommendations are made available to improve the design of the subsequent aid schemes. During this process, new ideas and proposals will emerge, regarding the methodology or other modalities of evaluation. This "*learning by doing*" will necessarily lead to adjustments and fine-tunings which will need to be reflected in subsequent versions of this document. In the medium to long term, evaluation could gradually lead to a "*culture change*" in the field of State aid to the benefit of both taxpayers and State aid enforcement.

2 General context and rationale

The State Aid Modernisation initiative[6] aims at re-focusing the Commission's enforcement efforts on larger aid schemes that are likely to have the biggest impact on the common market. In parallel, the analysis of cases of a more local nature and with little effect on trade will be simplified, *inter alia* by providing more flexibility for Member States in implementing such aid measures through an increased scope of general block exemptions. In order to maintain the overall balance and prevent undue distortions, greater simplification should be combined with greater transparency, effective evaluation and control of compliance with the State aid rules at national and European level.

State aid evaluation complements the toolkit available to assess and approve State aid. By providing better insight on the actual impact of aid schemes on trade and competition over time, it allows lightening the ex-ante assessment and its conditionality. It enlarges the possibility for more aid to be put in place without being subject to the Commission's ex-ante assessment, on the premise that possible undue distortions can be observed and if necessary corrected. State aid evaluation should remain within the legal framework of the Treaty. It should not ex-post put into question measures that would have been reasonably declared compatible ex-ante, following a proper assessment. It should also remain a proportionate exercise. It should therefore focus on a few aid schemes with the largest potential distortion on the internal market and not create an administrative burden on the granting authorities or on the beneficiary companies which is not proportionate to the benefits that it may generate.

In the specific context of EU State aid control, the overall objective of State aid **ex-post evaluation** is the verification of the balance between the public objective of the aid and its impact on competition and trade between Member States. State aid evaluation can explain whether and to what extent the original objectives of an aid measure have been fulfilled (i.e. assessing the positive effects) and to determine the impact on markets and competition (i.e. possible negative effects).

State aid evaluation should in particular allow:

[6] As laid down in the Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions, EU State Aid Modernisation (SAM), 8.5.2012, COM(2012) 209 final.

- Assessing whether the scheme is effective in achieving the *direct* objective for which it was introduced (e.g. whether a certain aid has effectively led to more investment on the part of the aid beneficiaries);
- Assessing whether the scheme provides for *indirect* effects on the objective of interest (e.g. whether the aid led to positive spill-over effects on other firms, or led competitors to lower their own investment effort);
- Detecting negative indirect effects, in particular the potential aggregated effect of large schemes on competition and trade (e.g. whether the aid led to crowding out or shrinkage of competitors or just displaced activity away from other locations, against the EU interest).

Based on the above assessment, evaluation can confirm whether the assumptions underlying the ex-ante approval of the aid scheme are still valid and/or help to improve the design of the scheme, introduce corrective measures, and calibrate State interventions to improve the effectiveness and efficiency of the aid to the extent necessary to guarantee that the overall balance between positive effects and distortions is still acceptable. Ex-post evaluation hence differs, in its purpose, from the two ex-post exercises currently carried out by the Commission with regard to State aid schemes:

- The **monitoring** exercise of the Commission is a periodic review of the legality of a sample of State aid measures implemented by Member States. It is designed to ensure the proper implementation of the Commission decisions and compliance with the pertinent legal provisions (i.e. those embodied in the General Block Exemption Regulation). In this context, the Commission verifies if the ex-ante rules and conditions are complied with within a representative sample of cases.
- The data collection in the **annual reporting** by Member States has as primary objective to increase the transparency of State aid granted by Member States but also to provide solid statistics for policy or monitoring purposes. Annual reporting data provide information primarily in quantitative terms (for instance, to show towards which objectives State aid was oriented and with what level of budget). The results of the reports by Member States are used for the preparation of the State aid Scoreboard[7].

Box: Link with impact assessments carried out under EU regulations

As stipulated in EU Regulations, evaluations shall be carried out to improve the quality of the design and implementation of programmes, as well as to assess their effectiveness, efficiency and impact. Impact of programmes shall be evaluated in accordance with the mission of the respective funds in relation to the targets for the EU strategy for smart, sustainable and inclusive growth (especially with regard to the EU2020 headline targets and other relevant impact indicators), as well as in relation to macro-economic indicators, where appropriate (such as gross domestic product, unemployment, etc.).

[7] http://ec.europa.eu/competition/state_aid/studies_reports/studies_reports.html

The Guidance document on monitoring and evaluation of the ERDF, Cohesion funds or of the ESF for the Programming Period 2014-2020 set out in detail the impact evaluation requirements. The Guidance document suggests using counterfactual impact evaluations (to understand whether the intervention works) complemented with theory-based impact evaluations (i.e. to understand why and how an intervention works).

Counterfactual impact evaluations carried out in accordance with the referred Guidance documents are sufficient to satisfy the first objective of State aid evaluation, i.e. to estimate the direct impact of the aid at firm level. Nevertheless, in order to get insight on competition related matters, such evaluations should be complemented with evaluations focusing on the public intervention's impact on competition and trade in the EU.

As indicated in the Broadband Guidelines[8] and in the Regional aid Guidelines[9], State aid evaluation could be carried out with respect to schemes which potentially have a significant impact on the internal market. Therefore the following types of aid schemes would benefit from evaluation for the reasons explained below.

Box: Example of schemes which would benefit from evaluation

Large aid schemes

In line with the SAM Communication, the largest aid schemes could be subject to ex-post evaluation, since (1) such schemes can impact the single market most severely; (2) the largest efficiency gains can accrue due to their high budgets; (3) large schemes with many different types of beneficiaries can provide sufficient data for the evaluation.

Certain aid schemes may still worth evaluating if despite their size they do not entail any specific problematic aspect because they are routine cases, without likelihood of significant changes or when no serious distortions could arise, for instance aid schemes for outermost regions, employment aid for disabled persons or compensation of natural disasters could be considered as falling into this category.

Novel aid schemes

The definition of "novelty" may vary across aid instruments and Member States. Novelty could in principle be considered in terms of the nature of the aid measure (for instance, the aid is pursuing a new objective or experimentation with new instruments) or aid targeting nascent or emerging markets (where market developments are at a very early stage). The latter schemes have the potential to deeply and permanently shape industries. The scope for both benefits and distortions is therefore particularly wide. Such novelty could include for instance, the introduction of a new capacity mechanism in the energy sector, aid to new types of

[8] EU Guidelines for the application of state aid rules in relation to the rapid deployment of broadband networks. OJ C 25 of 26.01.2013.

[9] Guidelines on National Regional aid for 2007-2013. OJ C 54/13, 4.3.2006.

technologies or a novel type of support for renewable energy sources in the environmental aid context.

Possibility of significant changes during the life of a scheme impacting its implementation

The possibility of significant (market, technological or regulatory) changes during the life of an aid scheme and the need for its ex-post evaluation are to be assessed on a case-by-case basis. Such significant changes could include, for instance, the foreseen revision of an applicable market regulation, aid to fast-moving industries where the market environment and the available technologies are changing/ developing at a rapid pace. Even when these changes can be expected, reduction of the duration of the scheme and reliance on a new ex-ante assessment may not always be appropriate or possible. If so, where significant changes occur, there is a risk that public funding is not effectively used (for instance targeting a 'market failure' which will disappear) and that significant distortions arise vis-à-vis new market entrants vs. incumbent companies, or new technologies vs. legacy technologies. As illustrative examples, the revision of an existing regulatory framework (for instance in the electronic communication sector), the high fluctuation of the input or output prices (for instance in case of solar panels) or the launch of a new technology on the market (for instance availability of fourth generation mobile network for broadband services) could duly justify a revision of the original scheme in the light of an ex-post evaluation, so that the future schemes can take new market developments into account.

Other schemes benefitting from evaluation

The individual Guidelines in the different policy fields could identify certain aid schemes where an ex-post evaluation would be particularly relevant.

For instance, aid measures targeting sectors characterised by significant overcapacity or at a constant decline[10] could disrupt market forces by preventing inefficient firms to exit from the market or giving undue advantage to a selected number of market actors to the detriment of other firms (which might even result in the market exit of more efficient companies).

3 The design and the process of evaluation

A proper evaluation of the impact of State aid measures should be objective, rigorous, impartial and transparent.[11] It should be conducted on the basis of sound methodologies, by experts who have the adequate experience and methodological knowledge to engage in the exercise. In order to ensure meaningful evaluation reports, this methodological guidance intends to promote best practices and to ensure high quality and comparable evaluations

[10] Constant, non-cyclical decrease in terms of output, employment, proportion in external trade,...

[11] See for instance, European Commission's binding Evaluation Standards, OECD Evaluation Norms and Standards, United Nations' Evaluation Standards, World Bank: Independent Evaluation: Principles, Guidelines and Good Practice.

across Member States. It is based on the significant developments in econometric techniques aimed at estimating the causal effect of policies which have accumulated in the last two decades.

The first evaluation objective is to find out how the aid beneficiaries would have performed if no aid had been given to them. This task can be greatly facilitated if one properly anticipates and prepares for the evaluation. Firstly, it requires thinking about how to identify the causal impact of the measure. In particular, this means for instance identifying ways to measure the impact of the State aid when there are also other factors determining the observed outcome (e.g. the level of R&D activity performed by firms may be influenced by many other factors than just the aid). Secondly, it requires making sure that the relevant data will be available by the time the evaluation is to be carried out. Both requirements can be best dealt with at the stage of designing the policy. The best practice is to design the evaluation at the same time as the policy itself in order to avoid ex-post a failure to comply with the obligation to evaluate.

A good planning and rigorous implementation of the evaluation plan, for instance in terms of data gathering, will significantly reduce the work required for the evaluation and will ultimately increase the quality of the report.

Box: The 'evaluation plan'

A good evaluation plan should contain at least the following minimum set of elements:

1. Objective of the scheme, key assumptions, objective of evaluation, identifying the relevant result indicators
2. Methodology
3. Monitoring, data gathering and availability
4. Timeline
5. The body conducting the evaluation (or the selection principles)
6. Peer review
7. Use of the results

The individual steps of the evaluation plan are set out in greater detail in the subsections below.

3.1 Objective of the scheme, key assumptions, objective of evaluation, identifying the relevant result indicators

The evaluation of a State aid scheme aims at determining whether the scheme has achieved its goals, while limiting distortions. Therefore, evaluating a scheme first requires to explicitly

state the rationale for the implementation of the scheme as well as its expected benefits and its potential distortions.

The impacts of a public intervention can be distinguished and assessed according to the following three levels, and amounts to asking the following questions:

1. *Direct* impact of the aid at the level of beneficiaries:
    - Did the aid really change the behaviour of the aid beneficiaries? (incentive effect)
    - What were the expected beneficial effects? To what extent were they achieved?
    - Did the effect of aid differ significantly across beneficiaries?
    - How efficient was the measure? What was the relative (cost-) effectiveness of the different aid instruments used?
2. *Indirect* impact of the aid on the policy objective of interest
    - For instance, did R&D aid lead to positive spill-over effects on the R&D activity of other firms, or did it lead competitors to increase/lower their own R&D activity;
3. Effects on competition and trade
    - What negative effects of competition and trade can be identified? For instance, did the aid crowd out competitors or attract activity away from other locations?

The ex-post evaluation should, to the extent possible, aim at assessing the impact of the public intervention at all three levels, addressing the relevant questions in relation to the specific objective of the aid scheme. However, the direct impact of aid on the beneficiaries (the incentive effect) is typically the type of impact that can most robustly be measured. It is also this type of impact for which most evaluation methods have been developed so far. Ex-post evaluation of the incentive effect is furthermore of paramount importance as it can provide valuable insight into the types of indirect effects and distortions that can be expected from the aid. In particular, where the aid provides no incentive effect, it can readily be assumed that the aid is distortive in the sense that it provides the beneficiaries in question with windfall gains.

The above list of questions should lead to the choice of concrete **result indicators** to capture quantified information about results achieved by the State aid scheme**.** An indicative list of results indicators in different areas can be found in Annex II. Importantly, this list is non-exhaustive and is purely indicative. The result indicators will have to depend on the objective of the aid that is evaluated and the evaluation plan. The evaluation should devote attention to explaining why the chosen indicators are the most relevant for measuring the outcome of the aid.

An evaluation should also try to compare the effectiveness of different tools or channels to achieve the same aim, whenever possible. For instance, if there were other tools that were used (such as public loans instead of grants), their relative effectiveness should be compared.

Example: Evaluation questions for regional aid

In regional aid, the main objective is to foster the economic development of certain regions in order to further enhance economic cohesion in the EU.

The result indicators measuring the outcome of aid will depend on the objectives of the aid scheme. For example, if the aim is to boost employment in a region, then the evaluation must assess whether aided undertakings indeed created more new employment than they would have absent the aid. The assessment should also consider the cost effectiveness of the measure: the cost of each created job is a meaningful indicator in this context. Other outcome indicators may be relevant as well. For example, the evaluation could assess whether in the long term aid is fostering growth by increasing the productivity of the aided firms and the overall productivity of the concerned regions. One could further assess the magnitude of the spill-overs to other existing firms in the region. The spill-overs may be positive if as a result of aid to some firms, all invest more than otherwise would have been the case. The spill-overs may be negative if the new jobs created by the aid are filled with workers attracted from non-aided firms. The evaluation can also assess how the effect of aid differs for firms of different characteristics, such as firm size and international orientation.

Example: Evaluation questions for R&D&I aid

The evaluation of R&D&I aid schemes should assess whether company's R&D&I activity increases due to State aid compared to what would have been the case absent aid. R&D&I activity is usually measured based on increased use of R&D&I inputs or outputs. The former includes own expenditure on R&D&I or the number of staff employed in this domain. R&D&I outputs encompass among other things patents, new product introductions or novel processes adopted. The key question to assess is whether the aid results in additional activity that would not have taken place otherwise, or whether the aid just crowds out the beneficiaries' private R&D efforts.

Other useful evaluation questions are whether aid beneficiaries undertake more risky research (e.g. in new applications), whether aid helps firms which were not innovative before to start innovating, whether the aid is cost effective (e.g. by focussing on the cost per extra researcher, the cost per extra patent). The evaluation may, where possible, also aim at addressing positive and negative spill-overs of the aid. Positive spill-overs could for instance be dissemination of knowledge or solving coordination failures. These positive effects may manifest themselves in the increase in R&D&I activity of other firms, i.e. firms that are closely related either from a geographical or technological point of view. Negative spill-overs on the contrary could be pre-emption of R&D&I projects by aid beneficiaries to the detriment of non-aid beneficiaries. These negative effects would materialize in a reduction of the R&D&I effort of other firms and may offset the direct benefits of the aid for the beneficiary.

Example: Evaluation questions for Risk finance aid

For risk finance aid, the main objective is to improve the provision of finance to viable SMEs from their early-development up to their growth stages and, in a longer-run, to develop a competitive SME finance market in the EU, which should contribute to overall economic growth. An evaluation should first aim at assessing that aided firms are facing a very severe credit constrained situation in the absence of the aid, e.g. compared to other firms in the economy, and whether the aid leads to increased levels of risk capital provided in the economy. The evaluation could further assess indirect effects, e.g. whether aided firms invest more than comparable firms, that they create more employment and that they have higher mid-term growth.

Example: Evaluation questions for environmental subsidies

Evaluating the experiences made with policies for the support of renewable energy technologies is important to continuously improve the design of renewable policies. Measures to evaluate the effectiveness of renewable support policies may include among other things the reduction of CO2 emissions as a result of aid or the number of firms deciding to go beyond mandatory environmental standards. These benefits should be compared with the costs resulting from the support of renewable energies.

3.2 Methodology: finding the good comparison

The objective of State aid evaluation is to identify the causal impact of the policy itself, undistorted by other variables that may have an impact on the observed outcome, e.g. general macroeconomic conditions or firm heterogeneity (e.g. differences in firm size, firm location, financial means or management capabilities).

This causal impact is the difference between the outcome with the aid and the outcome in the absence of the aid. While the outcome in presence of the aid is observed for firms who receive the aid, the outcome in the absence of the aid is only measured for firms which do not receive aid. By definitionone cannot observe what the outcome would have been *without the aid* for the firms which *received the aid*. To estimate the effect of the aid on aid beneficiaries, it is thus necessary to construct this counterfactual. This requires finding the most comparable firm(s) or *control group*.

The quality of this control group is crucial for the validity of the evaluation.

Firms which receive aid may well be in a different situation from firms which do not receive aid. They might for instance face different local demand and supply conditions, be more credit constrained, be more or less efficient, etc. All these factors may have an impact on the performance or activity levels of the firms, both when they receive aid and when they do not. Comparing the performance of beneficiaries with that of non-beneficiaries is likely to reflect

this reality more than the effect of the policy itself. An evaluation of the policy can therefore not rely on the simple comparison between beneficiaries and non-beneficiaries, it needs to take into account the differences in (observed and non-observed) characteristics between the two groups of firms.

Regarding for instance regional aid, aid beneficiaries from regions with various regional handicaps (e.g. in the local product, labour or capital markets) typically perform worse than non-beneficiaries in richer regions. However, this by no means reflects the effect of the aid itself. The relevant question is whether they performed better than *they* would have absent the aid, not whether they performed better than non-beneficiaries in richer regions.

Similarly, the identification of the aid effect needs to take into account general industry trends. Even if beneficiaries of regional aid reduce employment, it may still be the case that the aid is effective. For example, when industry conditions are worsening and all firms reduce employment, aid beneficiaries might reduce employment less than in the counterfactual. This is illustrated below showing a negative trend in terms of employment for aided firms, both before and after the aid was given. Nevertheless, the trend is less negative after the aid. It is this difference that identifies (or "isolates") the positive influence of the aid.

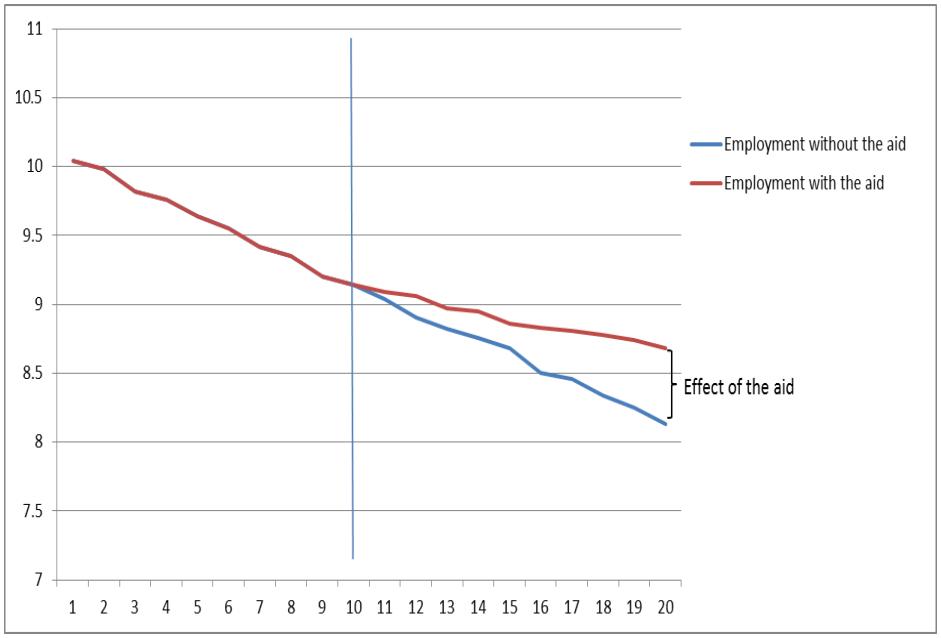

Figure 1 - Positive influence of the aid in negative trends

A specific identification problem emerges when non-beneficiaries decide themselves not to apply for aid. For instance, if all firms are eligible (i.e. all firms who come forward with a project and apply for aid get some aid), then the firms which do not apply may well be mostly those without a project. One might observe that firms that did not get aid performed worse in absolute and relative terms compared to the ones which did get aid. However, this finding might be entirely explained by the mere fact that the former group had no project to begin with, whereas the latter did. It is therefore crucial that firms in the control group who did not benefit from aid are part of that group for reasons that have no influence on the measured

outcomes. In particular, where firms have self-selected and voluntarily decided not to participate, this condition may not be fulfilled.

Any systematic difference between State aid beneficiaries and the non-beneficiaries has to be properly dealt with in the design of the evaluation in order to avoid the bias of the results (selection bias). Several reliable methodologies have been developed in the last decades to address this issue. The choice of the method depends on the details of the policy design and on the available data. Each of the methodologies has limitations and is only valid under a certain number of assumptions. Explicitly admitting and discussing these limitations and assumptions is crucial for the credibility of a study.

Randomization of the beneficiaries' selection process is one of the methods that allows for a proper evaluation. If aid beneficiaries are selected fully randomly, any systematic difference observed in the performance of the firms is due to the aid. However, this method is not always possible or easy to implement for practical or legal reasons. Other methods aim to use existing "accidental" or "exogenous" variations in the environment in which firms have to operate to identify causality. The most commonly used methodologies are Differences-in-Differences, Regression Discontinuity Design and Instrumental Variables. The technical appendix to this guidance paper presents in greater detail the most relevant methodologies, focusing on the most practical aspects and stressing the importance of a good approach towards identifying the causal effect.

We present in the boxes below some examples of such successful choices of comparison groups.

Example 1 (Regional aid): Criscuolo et al. (2012)[12] have evaluated the Regional Selective Assistance ("RSA") scheme in the UK between 1986 and 2004. In this period, RSA provided discretional grants to firms in disadvantaged areas. It was the main business support scheme in the UK. The scope for aid given under the RSA was governed by the Regional Aid Guidelines, in particular the maps of eligible regions ("regional aid maps"). These maps have evolved over time. By and large, the criterion for eligibility for a region is the relative position of the region in terms of GDP per capita or unemployment. Thereby, the status of a region can change either because it had developed over time or because the average EU per capita GDP changed (for instance when new member states joined the EU in 1995). Moreover, the indicators used to determine eligibility also change over time. Therefore, part of the change of eligibility of the firms does not depend on the situation of the firms themselves, but rather on events occurring outside of the UK or on changes in administrative rules. By focusing on this part of the changes in eligibility and assessing how these changes resulted in changes in investment activity, employment and productivity, Criscuolo et al. (2012) are able to convincingly identify the impact of the aid.[13]

[12] Criscuolo, C., R. Martin, H. Overman and J. Van Reenen, 2012."The causal effects of an industrial policy," CEPR Discussion Papers 8818, C.E.P.R. Discussion Papers.

[13] Technically, Criscuolo et al. (2012) are using an instrumental approach, as presented in the technical appendix.

Example 2 (Enterprise support): Martini and Bondonio (2012)[14] have examined two cases of enterprise support – an investment grant available throughout Italy (Law 488) and various SME schemes in the region of Piemonte. The first evaluation is particularly interesting. It compares the firms which saw their aid application approved (i.e. the aid-beneficiaries) with comparable firms who saw their aid application rejected as the budget that was available for the aid had reached its limit. The use of rejected applicants in the evaluation is particularly useful to avoid the selection bias which typically arises if one were just to compare applicants with non-applicants. This group of firms had passed the first quality check, which means that they had a credible investment project. Therefore, they shared with the aid beneficiaries the same ambition to invest in a credible project. However, because of budgetary limits (rationing), they did not receive aid. The difference in performance between (just) successful applicants and (closely) rejected applicants provided a reliable estimate of the effect of aid.

Example 3 (Loan guarantees): Lelarge, Sraer and Thesmar (2010) evaluate the effects of a loan guarantee program in France. The "SOFARIS" program provides insurance to lenders against borrowers' risk of default through guarantees. Borrowers pay an insurance premium, but this premium is subsidized. Lelarge, Sraer and Thesmar (2010) explicitly describe the nature of the selection effects. First, firms with more profitable projects are more likely to accept to pay the fee associated to the guarantee. Second, program managers are likely to select socially desirable projects which might not otherwise get access to private funding. Overall, firms self-select into the program and selection also occurs at the granting phase. This is likely to affect the results of naïve evaluations, based for instance on classical linear regressions or comparisons with the most comparable firm.[15] However, the factual and institutional context of the program provides a source of identification of the effects of the policy. The program was set up in the late 1980s and was initially restricted to firms active in the manufacturing and business services industries. In 1995, the public endowment of the program was increased and new industries (construction, retail and wholesale trade, transportation, hotels and restaurants and personal services) became eligible. Lelarge, Sraer and Thesmar (2010) compare the newly eligible firms to the previously eligible firms to assess the effect of the program on various indicators, like debt, employment, capital growth, financial expenses and probability of bankruptcy. Firms in these two groups are likely to differ. However, firms should be affected by similar macroeconomic shocks and therefore,

[14] Report for DG REGIO. A. Martini, D. Bondonio: "*Counterfactual impact evaluation of cohesion policy: impact and cost effectiveness of investment subsidies in Italy*" (2012).

[15] This is an instance where matching techniques, here one-to-one nearer neighbour matching, is not a better way to solve selection problems than ordinary least squares. As explained in the technical appendix, matching techniques are not, in general, a way to solve the issue of selection effects in the absence of natural experiments.

their differences should not change over time, except for the expected effects of the policy itself.[16]

Finally, the impact of multiple aid, either from one scheme, from several schemes or an ad-hoc aid, should be controlled for. If non-beneficiaries in the given program also receive aid from other programs or beneficiaries from the given program receive aid also form other programs, the evaluation of the effects of that given aid program are likely to be distorted.

3.3 Monitoring, data gathering and availability

Consistent and sufficient data has to be collected both on the aid beneficiaries as well as on the control group. Identifying this data and securing access to data sources will be a part of the evaluation plan.

Proper monitoring, data gathering and processing are crucial for the quality of the evaluation. Therefore, as soon as the aid scheme is approved, a mechanism should be put in place to monitor the intervention and to collect and process the appropriate data.

Making sure that the necessary data on aid applicants and beneficiaries is collected is a crucial step in the design of the evaluation plan. Providing this data can be made part of the eligibility conditions for aid.

Apart from data on aid application (including rejected applicants, when available), the data sources for aid beneficiaries and the control group must be identical, in order to be comparable, The necessary data will very likely involve combining multiple sources, such as databases containing information about aid receipts with firm registries. The evaluation may need to draw on existing data sources, such as administrative data (data from the tax office, data from the firm registry, data generated by community innovation surveys, patent office data, etc.). The evaluation plan needs thus to review the existing data sources, decide whether they are sufficient for the purpose of the evaluation and ensure timely access to them.

Data coming from administrative sources, for instance, national statistical offices, is likely to be made available to evaluators only under conditions relating to privacy and confidentiality of business data. Whenever necessary, the granting authority has to make all the possible steps to facilitate the access to this data by the experts carrying out the evaluation.

When data from several sources are used, it is crucial that these data are collected in a format that allows their consistent matching. It may be necessary to find unique identifiers for observation units in each data set used. For example, the firm and plant identifiers need to be unique in all datasets; the addresses have to be collected in a format that allows geo-localisation, etc. The exact origin of the identifier differ between Member States, could have a fiscal origin (like VAT numbers) or be directly provided by statistical institutes (like SIREN and SIRET in France).

[16] In practice, the authors implement a Heckman selection model with an exclusion variable at firm level and a classical IV strategy at sector level. See the technical annex for more details on these methodologies.

If data used for the evaluation is confidential, confidentiality needs to be guaranteed throughout the process of the evaluation. Nevertheless, confidentiality does not extend to the results of the evaluation, which are not confidential. In particular, no confidentially clause can be included in the contract for the evaluation, apart from non-disclosure obligations of confidential data and obligations to comply with general provisions of national statistical law and statistical secrecy for the presentation of the results.

Also access to data for replication in the peer review process should be ensured.

The evaluation of State aid could be complemented by information derived from surveys of aid beneficiaries and/or interviews of scheme managers. This information must, however be carefully interpreted given its subjective nature and the bias arising from potentially strategic answering, but it can be a useful source of (complementary) information.

3.4 Timeframe to carry out the evaluation

An evaluation plan should provide information on the precise timeline of the evaluation, including the duration of the scheme, the collection of data, carrying out the evaluation, submission of the final report and consideration of the implications of the evaluation report. When a meaningful evaluation can be carried out on the respective scheme, the timeframe could vary according to the objective of the scheme, its complexity, and the type of instrument. It should thus be discussed and agreed on a case-by-case basis.

If the Member State plans to re-notify the scheme (see *Figure* below), in order to learn from the impact of the ending scheme and remedy any potential flaws in the new scheme, evaluation of the original scheme should be carried out before the scheme ends.

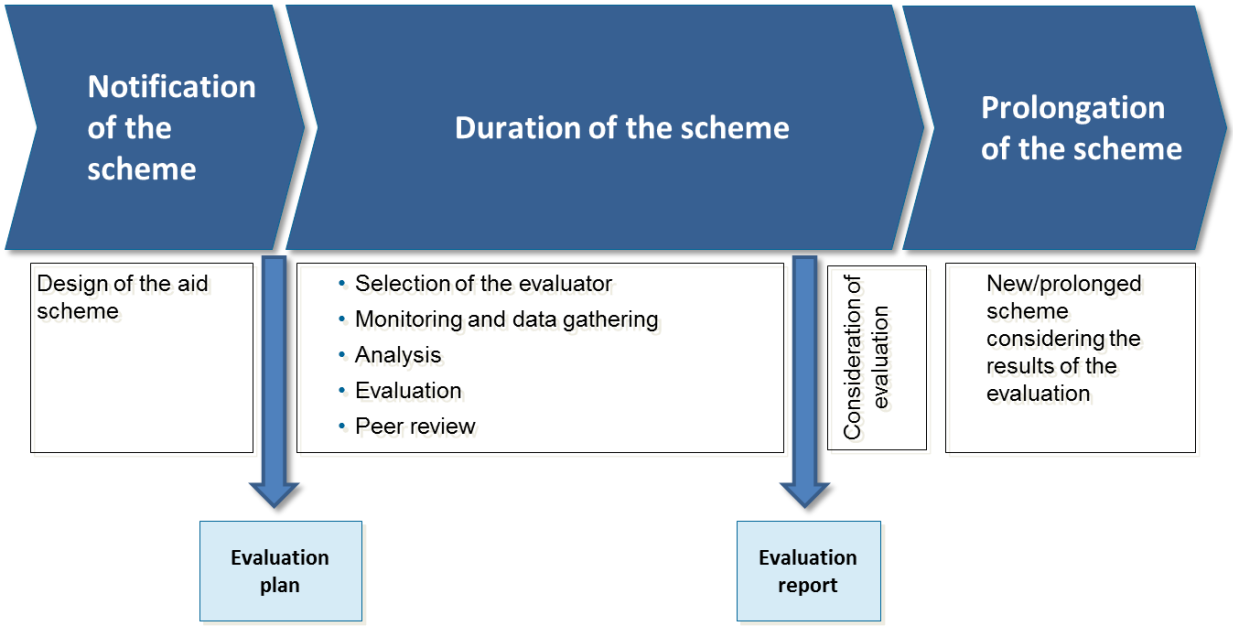

Figure 2 – Overview on the timeframe

3.5 The body conducting the evaluation

Best practice examples suggest that evaluations should be carried out by a body that is at least *functionally independent* from the granting authority and that has the necessary and proven skills and qualifications to carry out such evaluations. The functional independence of the evaluator from the granting authority is an important quality safeguard and contributes to the credibility of the evaluation.

3.6 Peer review and quality control

A further important tool to ensure the quality of evaluations is peer-review.

Peer review may be appropriate at crucial moments of the evaluation, i.e. at the stage of setting up the design of the evaluation and at the stage of actual evaluation. The goal of the peer review is to particularly assess the quality of the design of the evaluation, the plan to obtain data, the relevance of the methodology and the independence and level of experience of the scientific team carrying out the evaluation.

The aim of establishing such peer review is to obtain an expert report assessing the overall soundness of the evaluation (and not doing a separate evaluation instead), as well as to highlight its potential shortcomings. The identity or the selection principles of the peer-reviewers as well as their exact tasks would then have to be described in the evaluation plan.

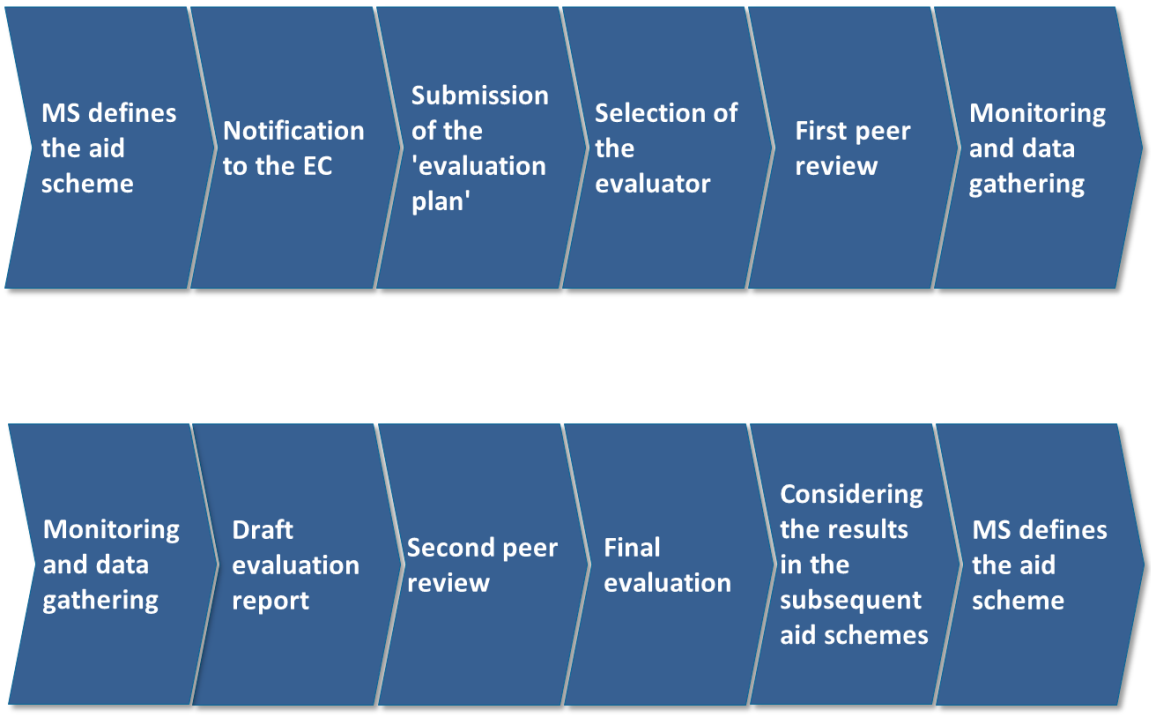

Figure 3 – Overview on the process of evaluation

3.7 Use of the results

The incorporation of the results of evaluation into the subsequent aid schemes is of pivotal importance.

3.7.1 Improving efficiency and effectiveness

The design of aid schemes are based on ex-ante assumption. However, the impact of the aid on markets may differ from expectations, for instance due to imprecise or incomplete assumptions or to unforeseen changes in the market conditions. Moreover, the impact of novel schemes may be difficult to anticipate. Based on the ex-post assessment, evaluation can help improving the design of the similar schemes in the future, introduce corrective measures and calibrate interventions to maximise effectiveness and efficiency, to the extent necessary to ensure that the positive effects of the scheme outweighs its distortions. Such improvements could vary from adjustments in the project design (e.g. change in selection criteria, reinforced checks on incentive effect, conditions to apply for aid) to more significant options (e.g. promoting the use of an alternative aid instrument, redefined objectives, redefined target beneficiaries, considering non-aid options to achieve the same policy objectives, including adapting regulation).

3.7.2 Support the ex-ante appraisal of the (re-)notification

Positive evaluation conclusions can facilitate the prolongation of the scheme and the ex-ante assessment of new schemes. In assessing whether a notified aid measure can be considered compatible with the internal market, the Commission analyses whether the design of the measure ensures that the positive impact of the aid in reaching an objective of common interest outweighs its potential negative effects on trade and competition. Evaluation may demonstrate whether and to what extent an aid measure meets the criteria for compatibility, thereby allowing for its prolongation or introducing useful feedback for its modification.

Better ex-ante assessments could also open the way for an enlargement of the set of measures that can be exempted from notification or subject to a lighter scrutiny in the future.

3.7.3 Advocacy of best practices

Advocacy of best practices from Member States would be useful both for the choice of the evaluation methodology and for the results of effectiveness and efficiency of different types of aid. Successful experience from other Member States could be used ex-ante to design the future aid measures more effectively. To promote best practices and thereby to promote the "good aid", workshops on effectiveness and efficiency could be organised by the Commission (on topics such as "How to design good environmental measures?"). Academics and other stakeholders should be involved and ex-post studies could be used as a starting point for discussion.

Annex I: Methodological appendix on causal inference

The objective of the evaluation is to measure the impact of the policy on the relevant outcome(s). It is to serve as a critical, evidence-based judgement of whether the intervention – i.e. the granting of State aid - has met the needs it aimed to satisfy and actually achieved its expected effects[17]. Many parameters and circumstances are likely to affect this outcome. The aim is to be able to delineate precisely what is the causal impact of the policy itself: how much of the observed changes in the outcomes can be attributed to having received the State aid?

A State aid scheme can have impact at very different levels. It is normally expected to have a direct effect at the level of the beneficiary. Understanding the magnitude of this effect is crucial to assess the level of efficiency and effectiveness of a public measure. However, since aid is directed towards firms which interact in markets or regions which compete to attract economic activity, State aid also normally has indirect effects. These effects could for instance be spill-over effects on other firms (e.g. positive spill-overs from R&D or the crowding out of investment by other competing firms) or displacement effects (e.g. shifts in economic activity from one region to another). These indirect effects are the basis for both the potential harm and the benefits stemming from State intervention in the economy. Therefore, evaluating public measures requires assessing the magnitude of these indirect effects as well.

Measuring the direct and indirect effects of a policy normally requires the use of different tools. The last decades have seen an important development of methodologies and techniques intended at assessing the direct effect of policies on their beneficiaries. These techniques are presented in greater detail later in this section. Unfortunately, it is only in rare circumstances that these techniques will also allow assessing the indirect effects of the aid measure on firms or regions. The evaluation of the indirect effects of the State aid scheme usually requires other types of evidence than what is used for assessing the direct effects on the recipients and interpretation normally relies more on economic theory and modelling. It is more difficult to provide precise guidance on this type of exercise as it has to be tailor made to the possible and expected positive and negative effects of the policy. Therefore, this evaluation has to be carried out after a careful and rigorous analysis of the most credible possible indirect effects of the policy. Based on this analysis, evaluators can derive measures based on micro data from non-aid beneficiaries, in particular in the same region, cluster or industry, as well as in neighbouring regions. This should form the core of the indirect assessment of the indirect effects of the State aid measures. If necessary, this can be complemented by more macroeconomic data and, most importantly, carefully chosen case studies.

The evaluation of direct effects is a necessary and crucial first step. However, a rigorously performed assessment of the indirect effects of the aid serves as an important piece of evidence in the assessment of the broader effects of the scheme. If the absence of additional

[17] Commission Guidelines for Evaluation, Working Draft of July 2013.

investment by aid beneficiaries is, broadly, indicative of failure of the policy, even a positive effect is not sufficient to conclude that a policy has fulfilled its objectives. In particular, if it turns out that the direct impact of the aid on the beneficiaries is very small or even non-existent, the scheme is very likely to be considered as not fulfilling its goal, unless very convincing arguments can be made about the existence of large and beneficial indirect effects. The contrary is also true: even if the evaluation finds that positive direct effects for the aid, the question remains whether there may be negative indirect effects that offset or even outweigh these.

Moreover, it is not always easy to clearly separate direct and indirect effects. A firm might have invested more (alleged direct effect) because its own investment has crowed out investment by competing firms (interacting indirect effect). A firm might also invest more because it expects spill-overs and investments by other firms. Moreover, it might be the aid itself or simply the granting of the aid which could have either effects. The likely presence, direction and expected magnitude of indirect effects should be discussed in detail in the evaluation of the direct effects. The economic theory that links the indirect effects to the aid should be explicitly Stated and additional information that may serve as evidence supporting this theory should form an integral part of the evaluation.[18]

Causal Inference

The causal impact of aid is the difference between the outcome with the aid and the outcome in the absence of the aid. The outcome in the presence of the aid is observed for firms who receive the aid. However, the outcome in the absence of the aid is only measured for firms who do not receive aid. By definition, we do not observe what the outcome would have been without the aid for the firms who received the aid. To estimate the effect of the aid on aid beneficiaries, it is thus necessary to construct this ***counterfactual***, i.e. to establish a reasonable scenario capturing what would have likely happened to the recipients of aid had they not received it. This requires finding a *control group*, i.e. a group of firms which should be as similar as possible to the group of firms that received the aid in all respects except for the aid itself.

The quality of the control group is crucial for the validity of the evaluation. Firms who receive aid typically differ in their characteristics from those who do not receive aid. They might for instance be active in a poorer area with less market potential, be more credit constrained, be more or less efficient, have a project to carry out or not, etc. Hence, naively comparing beneficiaries with non-beneficiaries is likely to reflect this reality more than the effect of the policy itself.

Making sure that this systematic difference between State aid beneficiaries and the non-beneficiaries (the so-called selection effect) does not bias the results is the core issue to carry out a valid evaluation. Several reliable methodologies have been developed in the last decades

[18] Although this document focuses on the direct effects of aid, the fact that the aid may have indirect effects does impose some analytical challenges on the assessment of direct effects, and special care has to be taken to the effects of market interactions.

to address this issue. The choice of the method depends on the policy to be evaluated and available data. In addition, each of the methodologies has limitations and is only valid under a certain number of assumptions. The credibility of a study can be increased by explicitly identifying and discussing these limitations. This technical annex presents the most relevant methodologies, focusing on the most practical aspects and stressing the importance of a good identification strategy.[19]

A. Randomised experiments

The identification of a proper control group is key to obtaining good (i.e. unbiased) estimates of the effect of the policy. The most favourable case is when there is no selection effect because beneficiaries were selected randomly[20]. Then, there is no systematic difference between beneficiaries and non-beneficiaries apart from the aid and the difference in the outcomes can be attributed to the policy.

However, random selection of aid beneficiaries is sometimes criticised for being at odds with the aim of many schemes to select the best possible aid beneficiaries on the basis of objective criteria. Still, in certain circumstances it might be possible to introduce elements of randomness in the eligibility or in the incentives to participate of beneficiaries. One example is setting a fixed budget for the given scheme. If the applicants' demand for support exceeds the budget and they are fairly equal in their characteristics, then one may try to establish randomness in treatment. Another example is randomly exposing potential recipients of aid to different levels of information about the scheme.

Pilot projects provide further opportunities for random allocation of aid. In case of innovative policies it might be advisable to evaluate a smaller scale pilot first. This pilot could have a smaller size and beneficiaries may more easily be chosen randomly. Another alternative would be to ramp-up a scheme, for instance to make eligible 25% of randomly selected firms the first year to respectively 50, 75 and 100% the second, third and fourth year (or alternatively, to advertise the scheme to a larger and larger audience). For a new policy, a period of ramp-up is in many cases an administrative necessity.

These ideas may be better suited for the implementation of totally new schemes or a large variation of existing schemes. It is probably fairly difficult to randomise eligibility (directly or indirectly) for the continuation of an existing scheme. However, this does not mean that random experiments cannot be used for parts of their evaluation. In particular, it is still possible to randomly select beneficiaries for potentially more efficient, more targeted and/or less distortive variants of the scheme. For instance, in the case of a grant scheme, it may be possible to randomly propose a newly introduced loan scheme instead.

---

[19] This annex offers a quick and non-technical presentation of the econometric methods for policy evaluation. This presentation takes many elements from Givord (2010), other very good presentations can be found in Imbens and Wooldridge (2009) and Angrist and Pischke (2008).

[20] Randomized experiments have for instance been the only acceptable methodology for the assessment of the effects of drugs and medical treatments for decades.

B. Quasi-experimental methods

Even though random experiments are the best possible way to evaluate the effect of policies, it is not always possible to implement them. Other methods have been developed to evaluate the effects of a policy from an ex-post perspective. They share the aim to use exogenous variations of the environment in which firms operate, to create situations very close to experiments (so-called natural or quasi- experiments).

It is generally a challenge for ex-post assessment to identify natural or quasi-experiments. However, a careful analysis of the design of the policy can entail an analysis of the existence of sufficient exogenous variation. If necessary, the initial setup can be adjusted to introduce more elements to allow identification of the effects of the policy.

a. Controlling for observable differences

As explained above, there normally exist significant differences between aid beneficiaries and non-aid beneficiaries. It is then necessary to account for these differences when comparing the outcomes between the two groups of aid and non-aid beneficiaries.

Many of the differences in characteristics are typically observable. The most common way to take these differences into account is to use linear regression. Linear regression seeks to control for the influence of observed characteristics on the outcomes. It assumes a linear relationship between the outcome, for instance the investment in R&D, and other characteristics of the firm, for instance the sector, age, size etc., including the granting of the aid. It is possible to see linear regression as a linear approximation of more complicated relationships[21]. Linear regressions can be seen as general purpose techniques and are used in many different evaluation contexts.

An alternative to linear regression is to use matching techniques. Matching techniques aim at pairing each beneficiary with another firm that "looks" very similar but did not receive aid. The observables used for matching can be firm characteristics or the estimated probability to receive aid (propensity score matching). Matching can be a useful way to control for observables in the context of a valid empirical strategy.

However, both simple linear regression and matching have some intrinsic limitations. Both are only valid under a conditional independence assumption. This condition requires that, once the impact of the observable characteristics has been taken into account, the outcome is independent of the observable characteristics. In practice, this normally requires that every variable that impacts both the outcome and the selection is observable (and is taken into account with the proper functional form). If this is not the case, both linear regression and matching will fail to provide a valid evaluation. For instance, if a firm has a "promising project", this both affects the likelihood that it will apply for aid (and get aid) and the

[21] Moreover, it is possible to interact characteristics (for instance sales and sector) and to introduce functions of these characteristics (for instance squares of variables).

likelihood that the firm is successful in growing a business. Not taking this into account will bias the results.

In particular, in the case of matching, comparing the outcomes between a beneficiary and its matched "twin" without aid, allows avoiding the selection effect only if the granting of the aid is unrelated to unobserved variables that also influence the outcome. In reality, this assumption will rarely be fulfilled. Measuring all the variables that have an impact on the fact of applying or getting aid is rarely possible. Implementing matching techniques moreover requires that firms who get aid are very similar in their observable characteristics to those not getting it. If the matched firms are truly similar in every observable aspect, the reason why some firms received aid and some others did not are, by definition, unobserved. The justification to the validity of a matching-based evaluation or a simpler classical linear regression can thus not be the mere existence of a very complete dataset with many observed characteristics.

On the contrary, the potential justification for the use of matching or simpler linear regression in evaluation relies on the fact that these unobserved reasons that explain eligibility or attribution of aid have no direct or indirect influence on the outcomes (once controlled for the observables). For an evaluation based on simple matching or linear regression to be valid, one would need to be confident that the set of firms who did not receive aid has been exogenously determined. This requires that once the observables are controlled for, there remains no unobserved factor explaining eligibility or attribution of aid that would also directly or indirectly influence the outcomes. In general, matching firms that are equally eligible for aid will not fulfil this latter criterion. For instance, if all firms are eligible, firms who get investment aid are much more likely to have a project than firms who did not get aid (as they would also have applied and been granted aid otherwise). Overall, firms with a project are more likely to grow in terms of sales or employment, but this is not related to aid and matching on observables is not able to disentangle the two (unless we measure the existence of a comparable investment project).

In many situations, the conditional independence assumption is bound to fail. It may therefore be necessary to implement different techniques than mere linear regression of matching to account for the existence of unobserved selection into the treatment.

The remainder of this section presents the most common methodologies used to assess policy impact in this context in more detail, i.e. Differences-in-Differences, Regression Discontinuity Design (RDD), Instrumental Variables (IV). These methodologies derive their validities from different assumptions and the best choice is normally driven by the context of the policy and the availability of data. This presentation sets out the merits and weaknesses of each particular technique. With the noticeable exception of randomized controlled treatments (“RCT”) presented above, there exists no technique superior to all the other ones in every aspects. The choice of a particular technique has to be guided by a careful analysis of the context of the measure and the available data.

It is worth stressing here that it is not the use of a specific econometric technique that allows identifying the effects of a policy; it is the exogeneity of the control group and hence the

quality of the counterfactual. The quality of the evaluation study will therefore crucially depend on how convincingly the researcher can establish the exogeneity of the control group. In cases where residual biases might remain, it is essential to discuss these biases in detail, including their sources and the directions and likely magnitude of their effects on the results.

b. Difference in Difference

Rationale and identification

As explained earlier, a simple comparison between beneficiaries and even a well-chosen group of non-beneficiaries is unlikely to lead to a valid evaluation. The reason for this is that it is not possible to exclude the existence of unobserved differences between the two groups, leading to a persistent difference in outcomes even in the absence of the aid. Moreover, simply comparing the outcomes before and after the aid for beneficiaries is also likely to lead to a spurious evaluation. It does not allow disentangling the effects of the aid from the effects of other factors that also affect the outcome, for instance the general economic trend, changes in the regulatory environment or increasing labour cost.

However, combining the two approaches might allow assessing the causal effect of the aid: this is the Difference-in-Difference approach. The general idea is to consider the difference in outcome between firms over time. Pre-existing differences would be attributed to other factors than the State aid. Only the change in these differences (the "Difference-in-Difference") would be attributed to the aid. In other words, the method compares the difference in the performance between beneficiaries and control group *before* the aid as well as *after* the aid and then attributes the change in the difference to the aid. The method works if, over time, both the beneficiaries and the control group are affected by the other factors that also affect performance in the same way. It can then be concluded that the aid is the only relevant factor that explains the observed change in performance of beneficiaries relative to the control group.

The crucial assumption is that the differences between beneficiaries and control group are stable over time and that both groups are affected identically by common shocks during the period. This assumption can fail in practice. For instance, if beneficiaries are the more vulnerable firms, they are likely to be more affected by economic downturns and general business climate. Therefore, the control group has to be made of as vulnerable firms. Overall, the choice of the control group is the key for the validity of the method. Identification does not lie in the use of differences-in-differences, which is the mere technical implementation, but in the proper choice of control group.

A special care in the construction of the control group is needed if non-beneficiaries decided themselves not to apply for aid. Applying or not for aid can be expected to be related to the returns of getting the aid. Therefore, there are reasons to believe that the anticipated outcomes of firms who do not apply for aid (in terms of employment, productivity, sales, etc.) differ from the expected outcomes for beneficiaries. For instance, if all firms who apply for aid get some aid, the only eligible firms who do not apply are those without a project. These firms are

not only likely to perform worse in absolute terms but also comparatively worse as time passes, while better firms implement projects and grow. Employment, productivity or sales cannot be expected to remain parallel and double differentiation does not, in general, solve the problem.

Therefore, firms in the control group who did not benefit from aid need to have been selected for reasons that have no influence on the measured outcomes. They cannot have self-selected and voluntarily decided not to participate. The most convincing setup is when non-participation is related to non-eligibility that is the consequence of a natural experiment. In this case, non-eligibility is unlikely to be due to unobserved factors that also have an influence on the outcomes. Control groups could for instance be firms located in regions no longer eligible for aid (if this eligibility is not related to their own performance but rather to an exogenous event).

Implementation

From a technical point of view, difference-in-difference methods can be implemented either within a linear regression model or with matching. In the former case, the control group is chosen independently of the observable characteristics and it is therefore overall comparable to the whole group of the aid beneficiaries. Then, observable differences are taken into account in a classical linear regression. In the second case, the control group is made of firms that are individually comparable to each aided-firm in the sample based on observable factors. The outcome for each firm is compared to the outcome of its most comparable firm(s) and the results are aggregated. The two methods are two different ways to take observable differences into consideration but there is no a fundamental difference in terms of identification of the causal effect of the policy.

Depending on the circumstances, it may be worthwhile to compare the variations of outcomes of the beneficiaries and the control group before the aid. If the outcomes systematically start diverging already before the aid has actually been granted, it is likely that the control group and the group of the beneficiaries are diverging for reasons unrelated to the aid and the method does not give a valid estimate of the causal effect of the aid. This does not constitute a rigorous test of the validity of the assumption: such a test does not exist. However, this is at least a useful first sanity check.

Additional methods and robustness tests can be used when there several potential control groups exist which are a priori valid. The first and most natural robustness check is to implement several difference-in-difference estimators and to compare the results. In addition, it is also possible to use these different control groups to build a more reliable estimate. Imagine a scheme targeted at SMEs in a particular region. Two potential control groups are the non-SME firms in this region or SMEs is an adjacent region. None of these firms voluntarily decided not to apply for aid, they were simply not eligible. Nevertheless, neither of these control groups is perfect: larger firms in the same region are likely to be affected differently by general economic trends while SMEs in an adjacent region might be subject to different regional shocks. Instead of choosing between these two possible difference-in-difference estimators, it is possible to combine them and implement a triple difference

estimator (DDD): starting from the 'classical' difference-in-difference between SMEs and non-SMEs in the concerned region, one can subtract the same difference-in-difference from the adjacent region to cancel the variation in outcomes between SMEs and non-SMEs in the region with aid. Alternatively, one could systematically try to build a synthetic control group, made of SMEs from several adjacent regions and non-SMEs from the same region in order better replicate the pattern of the outcome for the beneficiaries before the aid (see Abadie, Diamond and Hainmuller, 2010 for details).

Inference

In addition to a careful design and choice of control group, the issue of inference has to be specifically addressed. The notion of inference in this context refers to the question of whether the effects that have been estimated are really significant. Statistical significance is a different issue from economic significance. The second one refers to the magnitude of the estimated effects compared to the other relevant parameters from an economic theory point of view. As explained earlier, economic significance is crucial. However, this discussion is in principle only relevant when, from a statistical point of view, the effects are estimated precisely enough, i.e. one can exclude that there is no effect at all.

There are reasons to believe that a straightforward inference under standard assumptions (such as the homoscedasticity assumption and the assumption of no autocorrelation) is likely to overestimate the statistical significance of the effects.[22]

The first problem is related to clustering of data. If the control group as well as the group of the beneficiaries are very homogenous (even if distinct from each other), all firms in each group are likely to be affected by similar deviations from the mean (shocks). In statistical terms, this means that the error term has a common component. If the variance of this common component is large compared to the true idiosyncratic shock for each firm, the inference will be biased. With two periods and two groups, the problem can be particularly severe and borderline to an identification issue: it is impossible to separate the effect of the shocks shared within each group from the effect of the policy. The problem does not need to be as severe if the groups are not so homogeneous. However, it is always necessary to reflect on the presence of common shocks for homogeneous subsets of the groups. For instance if demand is local, it will normally be necessary to correct for the clustered structure of the error term at the level of localities. The same could apply to industries or sectors.

The second problem emerges when panel data are used. Error terms of most firm level data like employment, productivity and investment are normally auto-correlated. This means that deviations from the mean in one period are likely to persist in the next period. Ignoring this issue leads to overestimating the precision of the estimation of the effects and reject more often that it should the absence of effects of the policy. This problem can be severe, as shown in Bertrand, Duflo, and Mullainathan (2004).

[22] This issue has been emphasized largely in the context of difference-in-difference technique, but the same problems can emerge with the other techniques covered by this paper.

c. Instrumental Variables

Rationale and identification

Using instrumental variables ("IV") is a classical method to deal with endogeneity of explanatory variables. Since benefiting from aid can be seen as an endogenous explanatory variable of the performance of a firm in a linear regression context, it is natural to use instrumental variables to evaluate the effect of aid.

A variable is endogenous when it is correlated with an unobserved element, which also determines the outcome. For example, imagine that one tries to identify the effect of State grant on firms' employment by regressing employment on programme participation and other observables. Let us imagine that the aid programme targets underperforming firms who are likely to face difficult local market conditions. Market conditions are not observable by the evaluator and hence cannot be controlled for directly. However, when this variable is left unaccounted for, the effect of the grant is likely to be underestimated by the evaluator due to the endogeneity of programme participation. Whether the firm face favourable or difficult market conditions has an impact on both programme participation and on employment, i.e. programme participation is correlated with the error term explaining employment. The impact of market conditions on programme participation means that it is impossible to attribute the entire correlation between programme participation and employment to the causal impact of aid.

However, there also exist other factors explaining programme participation but not employment. For instance, as in Criscuolo et al (2012), geographical location may determine the total amount of money available for the programme in the region. Moreover the list of regions covered by the programme changes over time. If the programme budget for a given region changed over time for external reasons (e.g. average EU's GDP per capita dropped), this has an effect on the programme participation but not on firm's local market conditions. The change in employment that is related to the exogenous change in the programme coverage is not related to local market conditions. By focusing on this "part" of programme participation variable it is possible to isolate the true impact of the participation on firm's employment without interference of local market conditions. This is the logic of instrumental variables.

For the evaluation of State aid, an instrumental variable is a variable that can explain the fact of receiving the aid but has no direct impact on the other unobserved determinants of the outcome that has to be measured. Instrumental variables then allow focusing on the participation in the scheme without interference from the selection effects. For illustrative purposes, one can see the logics of instrumental variable as follows.[23] In a first step, programme participation is regressed on all the exogenous variables, including the instrumental variables. In a second step, the participation variable (the variable indicating whether the aid was received) is replaced with the participation as predicted in the first step:

[23] In practice, two stage least squares are implemented in one step for well-known inference reasons.

this expected participation is not correlated with the unobserved element that also determines the outcome.

Issues with weak instruments

An instrumental variable is a variable that can explain the fact of receiving the aid but has no direct impact on the other unobserved determinants of the outcome that has to be measured. This simple and classical definition, however hides a number of practical difficulties. There exist tests aimed at checking for the consistency of instruments when more instruments are used than what is strictly necessary to identify a model. However, there exists no test of the validity of instruments. The main focus of a study using instrumental variables generally is to explain why each individual instrument can be assumed to be uncorrelated with the unobserved determinants of the performance of the firms, would it be employment, productivity, sales, investment, etc. Such explanations, based both on economic arguments and factual elements, are necessary to assess the validity of the evaluation. However, they are not sufficient, especially when several instruments are used.

The discussion of the quality of instrumental variables should include the issue of weak instruments, i.e. instruments weakly correlated with the outcome variable. When instrumental variables are poorly correlated with the endogenous variable, estimates are likely to be imprecise. One might be tempted to add more instrumental variables in that situation. It is well known that by instrumenting with a large enough number of variables, it is possible to recover enough of the initial variable to get statistically significant results. At the same time, the two stage least squares naturally gets closer and closer to the biased ordinary least squares estimate.[24] The potential for such bias should be explicitly addressed in any evaluation using the IV method. In particular, the issue of the credibility, not only of their individual, but also of their joined exogeneity has to be addressed.

A special case is also when the endogenous variable is assumed to be auto-correlated. If the source of endogeneity is assumed to be solely contemporaneous, it is then possible to use past values as instrumental variables. However, one would then have to reflect on the exact validity of this approach. For instance, if explanatory variables are auto-correlated, this could also be the case of the measured outcome. Then, the lagged variables are also endogenous. More generally, if the autocorrelation of the explanatory variables is very large, exogeneity assumptions might fail. If it is small, one could resort to using many lags (and potentially future values) and would risk falling in the pitfall of using many weak instruments described before. Overall, instrumenting by past values could be a valid strategy but it should be used with the highest caution.

[24] A very interesting practical discussion about the biases created by weak instruments can be found in Bound, Jeager and Baker's (1995) discussion of the statistical biases in Angrist and Krueger (1991). Moreover, instrumental variable estimates are biased at finite distance. Therefore, even with sufficient datasets to ensure apparent statistical significance, non asymptotic biases can still be important.

Generally speaking, to avoid the problems described earlier, it is highly advisable to only use a small number of convincing instruments. It is then, however also necessary to show that the instruments are good predictors of the endogenous explanatory variable.[25]

Variations of two step estimations: Heckman (1979) selection model

The first regression of the two stage least squares can be seen as a linear probability model of the probability to be treated. This linear probability model is a linear approximation. However, in some cases, the probability to be treated, even restricting to eligible firms, might be low. Then, linear approximations might be too coarse to effectively approximate this probability to be treated and to focus on the tails of the distribution, which are precisely the matter of interest. There are several ways to deal with this issue. They all rely on replacing the linear probability model of the probability to be treated by a non-linear function.[26]

A classical approach is to treat the evaluation problem in the context of a selection model (Heckman, 1979). This approach treats the selection effect as an omitted variable problem in the linear regression of the outcome on the observables and the participation. [27] There exist several variants around this methodology, for instance by estimating the whole model by maximum likelihood, or instrumenting the granting of aid by the predicted value of the selection equation.[28]

However, it is crucial to reflect on the identification and in particular on the choice of variables. It is not satisfactory to use the same variables in both steps of the estimation, even if the results are sufficiently precise. [29]

It is only reliable to estimate a selection model with a so-called exclusion variable. An exclusion variable is a variable that explains selection of the aid beneficiary but not the

---

[25] This can take the form of computing the Fisher statistics of the first step regression. The higher this Fisher, the less likely it is that instruments are weak. Stock, Wright and Yogo (2002) propose a formal test. For one instrument, it is for instance necessary that the Fisher statistics of the first step regression is larger than 10.

[26] This section provides a very brief description of the selection models in this section. For a more complete presentation, the reader is referred to the seminal paper of Heckman (1979) and, mainly, Wooldridge (2002), chapter 17.

[27] This omitted variable is the difference in conditional expectation of the outcome for the selected sample (here the aid-beneficiaries). Under certain assumptions on the selection process of the aid beneficiaries (for instance a probit or logit model), this difference can be formally derived (the inverse Mills ratio) and is a function of the selection parameters. Then, the effect of the policy can be identified by adding the omitted variable to the regression. The selection parameters are unknown, but consistent parameters can be recovered in a first step estimation of the selection process. This leads to the estimation procedure sometimes referred to as "Heckit". It first requires recovering the parameters of interest for the selection of the aid beneficiaries, for instance a probit or logit specification. Then, a consistent estimator of the effect of the policy can be recovered by adding the estimated inverse Mills ratio to the linear regression. Statistical software packages normally have a feature to perform this Heckman estimation.

[28] For the presentation of all these methods, readers can for instance refer to, Wooldridge (2002), chapter 17.

[29] When the selection equation is non-linear, the inverse Mills ratio is not collinear to the other explanatory variables, even when the first equation includes only a subset of these explanatory variables. Then, in theory, the model is already identified. In this case, the inverse Mills ratio very often does not show enough variation, which leads to very imprecise estimates. However, especially with large samples, the estimation could still lead to significant results. Nevertheless, when all the variables of the selection model are also in the main equation, the model is solely identified due to the non-linearities of a particular parametric form.

outcome. It is not sufficient to remove one variable from the main equation to add it to the list of explanatory variables of the selection equation. On the contrary, this exclusion variable has to fundamentally explain the selection but have no impact on the outcome one is trying to explain. It is in substance very close to a valid instrumental variable. The choice of such a variable cannot be driven by convenience; it has to come from economic theory and/or experience.

d. Regression Discontinuity Design

Regression discontinuity design (RDD) is the latest addition to the classical evaluation toolbox.[30] It has known a large success in the academic community in the last decade, mostly due to its simplicity. This method exploits the existence of a variable which has a discontinuous impact on the probability to be affected by a policy. In the context of State aid schemes, several types of discontinuities can be useful. The first one is geographical borders: the eligibility of schemes can be linked to precise administrative borders, like localities, NUTS regions, etc. The second one comes from conditions imposed to the firms to benefit from a scheme, in particular in terms of age and size.

Let us consider an example. Some firms are formally excluded from the scheme even though they are very close to some firms who benefit from the scheme. Imagine for instance that projects presented by firms are rated by points (out of 100) and firms who get at least 70 points get aid while the others get no aid. A firm who scores 71 has a marginally better project than a firm who scores 69. However, the consequence of this marginal difference is dramatic: one gets some aid, while the second gets no aid at all. Comparing the outcomes for these two firms is thus very indicative of the causal effect of the aid.

Formally, the RDD requires that the probability to receive aid is discontinuous, while all the other variables are continuous.[31] The technical implementation can be very close to this of instrumental variable, using a threshold crossing dummy as instrument. However, there are two main differences. The first one is that RDD relies on weaker assumptions. In particular, we do not a priori require the independence of the instrument. For instance, in the case of the scoring, firms with better project might apply more than firms with bad projects. The only requirement is that around the threshold the probability to apply should not be discontinuous. The second difference is that the estimates are built only on firms very close to both sides of the threshold. Weaker assumptions thus come at a cost: RDD estimates are even more local than estimates by instrumental variables generally are. If the effects of the aid differ for firms

[30] A formal and complete description of RDD can be found in Imbens et Lemieux (2008).

[31] Formally, there are two different regression discontinuity designs: the sharp and the fuzzy design. In the sharp design, which is implicitly the one described here, all firms, and only them, are treated above a certain threshold. In the fuzzy design, the discontinuity is less drastic: there is a discontinuity of the probability to be treated, but this does not change from 0 to 1. In absolute term, as far as state aid schemes based on eligibility conditions are concerned, it is only if one considers the treatment to be the eligibility that the design is sharp. Otherwise, if the treatment is to receive aid, we are in a fuzzy design. On the contrary, when the allocation is based on a scoring, we only consider firms who apply and the design is sharp.

further away from the threshold, the RDD estimates are not a correct estimate of the effect on all aid beneficiaries.

The locality of these estimates can be of concern if one would expect large discrepancies of effects away from the discontinuity. Moreover, individual companies just on the other side of the border could be very significantly affected by the policy. This could for instance be the case if displacement effects are important. Then, the use of RDD at the geographical border is not a relevant empirical strategy. Last, the strength of the RDD is to focus on a narrow bandwidth around the discontinuity. If the bandwidth is large, the impact of the other characteristics cannot be assumed to be constant. The issue is normally not solved by controlling for the observables, which assumes a particular functional form.

Graphical inspection of the data can provide comfort as regards the reliability of the assumption underlying RDD. In particular, it is very important to control for three things. The first one is that there indeed is a discontinuity on the granting of aid. The second one is that the outcomes to be measured have a discontinuity at the same moment and no other discontinuity of the same kind anywhere else. Third, it is also necessary to check that there exists no discontinuity in the other parameters correlated with the outcome, including the propensity to apply for aid.

At last, discontinuities might be created deliberately in order to allow an evaluation of the scheme. In particular, ramp-up of policies could also be used to create discontinuities and help the identification of the effects of a policy.

C. Structural estimation

In some instances, it is possible to go a step further and confront a theoretical model, for instance of firms' investments, with the data in order to recover the key parameters of interest. This approach is qualitatively different from those presented before. Structural estimation uses a completely specified theoretical model of firm behaviour under all needed assumptions. Estimation then allows recovering parameters determining firm's behaviour in the theoretical model. This allows an evaluation at the closest of the determinants of the individual behaviour of firms and enables to carry out simulations about the efficiency of other tools. However, structural estimation is generally more demanding in terms of resources and data as well as in terms of assumptions.

It is impossible to provide general guidance on structural estimation as the identification, estimation and inference has to be derived on a case by case basis. Nevertheless, the general guidance provided before still applies. First, it is necessary that the theoretical model matches the key stylized facts of the market. Second, the issues of unobserved characteristics and selection have to be explicitly and properly addressed.

D. Additional methodological remarks

Heterogeneity of treatment effects

The previous sections focused on the estimation of the average treatment effect. The very name suggests that the effect of the aid varies between beneficiaries. This heterogeneity may have many roots and many consequences. The first consequence might be that, if aid is very effective for some firms but much less for others, the average effect might be statistically insignificant. This absence of statistically significant effect does not mean that the aid has no effect for any firms. From a policy perspective, the average performance of a scheme is a very interesting first indicator. However, trying to understand the determinants of this heterogeneity is as important for the design of better schemes. It allows focusing directly on firms where the aid is the most effective and least distortive.

Thereby, whenever possible, the effect of the aid should be estimated for different types of firms, such as small firms vs. large firms, young firms vs. old firms, innovative firms, credit constrained firms, etc. [32]

Distortions on the non-aided firms

Evaluating the impact of the scheme on non-participants, either directly or indirectly, is very informative for the evaluation of State aid. State aid may be distorting markets via effects on the non-beneficiaries, for example by knowledge spillovers from beneficiaries or by the reduction in relative competitiveness vis-à-vis beneficiaries, etc.

Moreover, the effects on the non-aided firms or locations can have an effect on the validity of the evaluation. For example, a part of the effect of regional aid could materialize by opportunities at the border: firms historically located on the "wrong" side of the border moving their location just on the other side. Then, an RDD at the border would mostly capture this displacement effect and would risk overestimating the real aggregate effect of the policy. In such a situation, another empirical strategy has to be used (for example it may be useful to check the robustness of the evaluation on wider regions).

E. Data

Having access to appropriate microeconomic data that enables conducting impact evaluation is crucial. These data have to be consistent between beneficiaries and non-beneficiaries. Therefore, they need to have the same source, with the natural exception of information on the aid itself. The data should be accessible at the most refined level although in some cases some form of aggregation at a later stage may be necessary.

[32] Another approach would be to systematically estimate different treatment effects for firms in different points of the conditional distribution. There is a growing body of literature estimating such quantile treatment effects, starting from Abadie, Angrist and Imbens (2002). This is a very useful tool to understand the intrinsic nature of the heterogeneity of treatment. However, it is less useful from a strict policy point of view, unless it is possible to directly target different firms depending on their position in the conditional distribution.

Data capturing the result indicators of both the treatment as well as the control group are necessary, including the time at which the outcome is measured. Furthermore, as much data as possible on factors potentially influencing outcomes and the entities' decision to participate in the aid programme are necessary. This data is used to 'control for' differences between the treatment and control groups. For example, on the firm level such data may include location, size and demographics, as well as production inputs used.

The most natural source of data is of administrative origin, such as fiscal balance sheet data, or large firms' national surveys. These sources provide information on the location and activity of firms, and sometimes of individual plants. They normally allow to track investment and sales by activity as well as to compute financial ratios. Large national or community surveys, such as Community Innovation Surveys are also of interest. They cover a large and representative sample and provide very complementary information on specific topics. Last, merged employer-employees datasets are also a relevant source of information. They normally allow relating labour characteristics to each plant location. This can be crucial when the geographical dimension of labour is a matter of interest.

Apart from indicators on results and recipient characteristics, data about the aid and the aid granting process is necessary. This information would usually come from the aid granting authority itself. This includes data on the amount and timing of granting of the aid to beneficiaries. However, general data on the process of attribution of the aid is also particularly helpful. Data on rejected applicants is important, especially if the granting of the aid is made using a scoring mechanism.[33]

Access to this confidential data is normally regulated. Securing in a timely manner access to this data for the whole of the scientific team performing the evaluation is therefore crucial. Moreover, these administrative sources are normally accessible at a delay. It is important to take into account data availability when designing the evaluation plan.

[33] Having data on rejected applications is particularly valuable for studies pursuing a regression discontinuity approach.

Annex II: List of possible results indicators

It shall be noted that below is an indicative list for illustration purposes only. The actual result indicators shall be set in accordance with the objective of the aid scheme and that of the evaluation.

Direct **impact of the aid at the level of beneficiaries**

	RESULT DIMENSION	RESULT INDICATORS
Regional aid	Positive impacts	Private investment matching public support Employment increase in the supported enterprises
Research, development and innovation aid	Additional RDI activity	Private investment matching public support Additional RDI expenditure undertaken by supported companies Number of new researchers employed in supported companies Number of new patents registered Number of enterprises supported to introduce new to the markets
Environmental aid	Positive environmental impacts	Reduced CO2 emissions of the beneficiary firms Additional capacity of renewable energy production Reduction of the share of waste landfilled or incinerated, Number of contaminated sites cleaned
	Early adoption of environmental standards	Percentage of companies reaching new environmental standards at least X months/years before they come into force
Energy (infrastructure) aid	Reduced energy consumption	Number of households with improved energy consumption classification Decrease of annual primary energy consumption of public buildings Number of additional energy users connected to smart grids
	Renewable energy support	Production share of energy from RES

Risk finance	Positive impacts	Returns achieved in the fund; Leverage of private investments Number of firms receiving risk capital
	Picking losers:	*poor average performance of investee firms due to deficient commercial management/insufficient private participation*
	Lack of sufficient degree of diversification	*Too small/regionally constrained funds with limited return prospects that remain unattractive for private investors*
Broadband aid	Increased broadband coverage	Additional household coverage with at least 30 Mbps broadband connection Additional household coverage or take up with at least 100 Mbps broadband connection
	Efficiency	Investments costs/aid per connecting a household (homes passed) Number of households signing up to new services
Rescue and restructuring	Positive impacts	Maintenance of employment and activity at firm-specific and regional level Changes in market share and productivity of aided firms
Aviation	Positive impacts	Number of air carriers using the airport; Private investment matching public support; Increase in regional productivity and/or gross value added (GVA)*)*
	Negative effects	*Duplication of lossmaking infrastructure or air routes;* *Deterioration of traffic of existing infrastructure (e. g. other airports in the catchment area or other means of transport*

Indirect impact of the aid on the policy objective of interest

	RESULT DIMENSION	RESULT INDICATORS
Positive effects	Macro-economic gains	Employment increase Increase in productivity and/or gross value added (GVA)
	Diversification of the regional economy	Number of industries under different NACE codes
	Increased cooperation between private and public	Number of enterprises cooperating with research institutions
	Positive externality / spill-over effects	Number of indirect beneficiaries (e.g. number of third parties accessing the facility) Changes in employment or activity in other firms and regions (aviation) Number of inhabitants with improved transportation means in the catchment area;

(Negative) effects on competition and trade

	RESULT DIMENSION	RESULT INDICATORS
Negative effects	Sectoral bias	Aid was predominantly granted to one industry in a multi-sectoral scheme
	Bias towards loss-making firms or firms with low productivity (prevention of exit)	Proportion of high vs. low productivity firms
	Bias towards incumbents	Proportion of old vs. young firms
	Reinforce the market power	Change in market power of a dominant player
	Location effect	Relocation from a poorer region to a more developed one
	For security of supply	Locking-in in high-carbon energy sources, Assess whether the concerns in terms of black-outs are real and continue to exist; Foreclosure of national electricity markets
	For energy infrastructure:	Foreclosure of national electricity markets, reinforce the market power of an incumbent

	Rescue and restructuring	Changes in employment or activity in other firms and regions Changes in market share and productivity of aided firms
	Aviation	Duplication of lossmaking infrastructure or air routes; Deterioration of traffic of existing infrastructure (e. g. other airports in the catchment area or other means of transport

Annex III: Glossary

Baseline	The value of the indicator before the policy intervention at stake is undertaken.
Control group	Counterfactual analysis requires finding the most comparable firm(s) or control group, i.e. a group of firms which should be as similar as possible to the group of firms that received the aid - except that they have not benefitted from that aid.
Counterfactual	To estimate the effect of the aid on aid beneficiaries, it is necessary to construct a 'counterfactual', i.e. to establish a reasonable scenario capturing what would have likely happened to the aid beneficeries if they had not received it.
Evaluation	The systematic collection and analysis of information about programmes and projects, their purpose and delivery; it derives knowledge on their impact as a basis for judgments. Evaluations are used to improve effectiveness and inform decisions about current and future programming.
Impact	The change that can be credibly attributed to an intervention. Same as "effect" of intervention or "contribution to change".
Indicator	A variable that provides quantitative or qualitative information on a phenomenon. It normally includes a value and a measurement unit.
Method	Methods are families of evaluation techniques and tools that fulfil different purposes. They usually consist of procedures and protocols that ensure systemisation and consistency in the way evaluations are undertaken. Methods may focus on the collection or analysis of information and data; may be quantitative or qualitative; and may attempt to describe, explain, predict or inform action. The choice of methods follows from the nature of the intervention, the evaluation questions being asked and the mode of enquiry – causal, exploratory, normative, etc.
Result	The specific dimension of the well-being of people that motivates policy action, i.e. that is expected to be modified by the interventions designed and implemented by a policy. Examples are: the mobility in an area; the competence in a given sector of activity.
Result indicator	An indicator describing a specific aspect of a result, a feature which can be measured. Examples are: the time needed to travel from W to Y at an average speed, as an aspect of mobility; the results of tests in a given topic, as an aspect of competence; the share of firms denied credit at any interest rate, as an aspect of banks' rationing.

Annex IV: References

Abadie, A., J. Angrist and G. W. Imbens (2002): "Instrumental Variables Estimates of the Effect of Subsidized Training on the Quantiles of Trainee Earnings," Econometrica, 70(1), 91–117.

Abadie, A., A. Diamond and J. Hainmueller (2007) : "Synthetic Control Methods for Comparative Case Studies : Estimating the Effect of California's Tobacco Control Program," Journal of the American Statistical Association, June 2010, Vol. 105, No. 490,

Angrist, J. and A. Krueger (1991) : "Does Compulsory School Attendance Affect Schooling and Earnings," Quaterly Journal of Economics, 106.

Angrist, J. and J. Pischke (2008), "Mostly Harmless Econometrics: An Empiricist's Companion", Princeton University Press.

Angrist, J. D., and J. Pischke. 2010. "The Credibility Revolution in Empirical Economics: How Better Research Design Is Taking the Con out of Econometrics." Journal of Economic Perspectives, 24(2): 3-30.

Bertrand, M., E. Duflo and S. Mullainathan (2004) : "How much should we trust differences-in-differences estimates ?," The Quarterly Journal of Economics, 119, 249–275.

Bound, J., D. Jeager and R. Baker (1995) : "Problems with Instrumental Variable Estimation When the Correlation Between the Instruments and the Endogenous is weak," Journal of the American Statistical Association, 90(430), 443–450.

Commission Guidelines for Evaluation, Working Draft of July 2013.

Criscuolo, C, R. Martin, H. Overman and J. Van Reenen(2012), "The causal effects of an industrial policy," CEPR Discussion Papers 8818, C.E.P.R. Discussion Papers.

Duflo, E., R. Glennerster & M. Kremer (2007), "Using Randomization in Development Economics Research: A Toolkit," CEPR Discussion Papers 6059, C.E.P.R. Discussion Papers.

Duflo, E., and M. Kremer (2005): Use of Randomization in the Evaluation of Development Effectiveness," in Evaluating Development Effectiveness, ed. by O. Feinstein, G. K. Ingram, and G. K. Pitman. New Brunswick, New Jersey and London, U.K.: Transaction Publishers, vol. 7, pp. 205{232}.

European Commission's binding Evaluation Standards. Available at: http://ec.europa.eu/dgs/secretariat_general/evaluation/docs/standards_c_2002_5267_final_en.pdf.

Garicano, L. C. Lelarge and J. Van Reenen, 2012."Firm Size Distortions and the Productivity Distribution: Evidence from France," CEP Discussion Papers dp1128, Centre for Economic Performance, LSE.

Givord, P. (2010) « Méthodes économétriques pour l'évaluation de politiques publiques », WPD3E n° G2010-08.

Givord, P., Rathelot, R. and P. Sillard (2013), "Place-based tax exemptions anddisplacement effects: An evaluation of the Zones Franches Urbaines program, Regional Science and Urban Economics", Volume 43, Issue 1, January 2013, Pages 151-163

Heckman, J. J. (1979), ''Sample Selection Bias as a Specification Error,'' Econometrica 47, 153–161.

Imbens, G. and J. Wooldridge (2009), *Recent Developments in the Econometrics of Program Evaluation*, Journal of Economic Literature, 47:1, 5-86

Imbens, G. W., et T. Lemieux (2008): "Regression discontinuity designs : A guide to practice," Journal of Econometrics, 142(2), 615–635.

C. Lelarge, D. Sraer and D. Thesmar (2010), "Entrepreneurship and Credit Constraints: Evidence from a French Loan Guarantee Program," NBER Chapters, in: International Differences in Entrepreneurship, pages 243-273, National Bureau of Economic Research, Inc.

Keane, M. P. (2010), "A Structural Perspective on the Experimentalist School." Journal of Economic Perspectives, 24(2): 47-58.

Martini, A. and D. Bondonio (2012), "Counterfactual impact evaluation of cohesion policy: impact and cost effectiveness of investment subsidies in Italy", Report for European Commission, DG Regio.

Nederlandse Rijksoverheid (2012), "Durf te meten", Eindrapport Expertwerkgroep Effectmeting, available at http://www.rijksoverheid.nl/documenten-en-publicaties/rapporten/2012/11/23/durf-te-meten-eindrapport-expertwerkgroep-effectmeting.html

Nevo, A. and M. D. Whinston. 2010. "Taking the Dogma out of Econometrics: Structural Modeling and Credible Inference." Journal of Economic Perspectives, 24(2): 69-82.

OECD Evaluation Norms and Standards. Available at: http://www.oecd.org/dac/evaluation/dcdndep/41612905.pdf.

Programming Period 2014-2020, Monitoring and Evaluation of European Cohesion Policy, European Social Fund, draft Guidance document of 9 January 2013. Available at: http://ec.europa.eu/social/main.jsp?catId=701&langId=en.

Proposal for a Regulation of the European Parliament and of the Council laying down common provisions on the European Regional Development Fund, the European Social Fund, the Cohesion Fund, the European Agricultural Fund for Rural Development and the European Maritime and Fisheries Fund covered by the Common Strategic Framework and laying down general provisions on the European Regional Development Fund, the European Social Fund and the Cohesion Fund and repealing Council Regulation (EC) No 1083/2006. COM (2011)

0615 f/2. Available on the internet at: http://ec.europa.eu/regional_policy/sources/docoffic/official/regulation/pdf/2014/proposals/regulation/general/general_proposal_en.pdf.

Sims, C. A. (2010). "But Economics Is Not an Experimental Science." Journal of Economic Perspectives, 24(2): 59-68.

Stock, J., J. Wrightand M. Yogo (2002): "A Survey of Weak Instruments and Weak Identification in Generalized Method of Moments," Journal of Business and Economic Statistics, 20(4), 518–29.

The Programming Period 2014-2020, Guidance document on monitoring and evaluation – European Regional Development Fund and Cohesion Fund – Concepts and Recommendations. April 2013.

United Nations' Evaluation Standards. Available on the internet at: http://www.uneval.org/papersandpubs/documentdetail.jsp?doc_id=22.

Wooldridge, J. (2002), *Econometric Analysis of Cross Section and Panel Data*, Cambridge: MIT Press.

World Bank: Independent Evaluation: Principles, Guidelines and Good Practice.