European
Commission

Evalsed Sourcebook:

Method and Techniques

Regional and Urban policies

Contents

Introduction

This Sourcebook describes a wide range of methods and techniques that are applied in the evaluation of socio-economic development. The methods and techniques are listed alphabetically, with two large sections on impact evaluation – theory based and counterfactual which discuss a number of approaches within the sections. Users are advised to search for material they want rather than reading through the sourcebook from the beginning to the end.

Choosing methods and techniques

The choice of methods and techniques stems from the evaluation design or mode of enquiry. Methods and techniques are selected if they are appropriate for answering the evaluation questions.

As elaborated in the GUIDE the choice of methods and techniques depends on:

- The type of the socio-economic intervention;
- The evaluation purpose - accountability, improving management, explaining what works and why, etc.; and
- The stage in the programme/policy cycle - prospective analysis/retrospective analysis.

Additionally, the appropriateness of the methods and techniques depends on the scope of the evaluation - which could range from an overall evaluation of a multi-sectoral programme, to an in-depth study of a particular evaluation question.

The elaborations of the techniques provide users with some ideas on how they can be applied and the main steps involved. It should be stressed, however, that some of the techniques are themselves longstanding and build upon a wealth of experience and literature that is not fully reviewed here. The main purpose of the presentations is to show how the techniques can contribute to the evaluation of socio economic development. Users are encouraged to refer to the references given prior to using the techniques for the first time. The information given here should however be sufficient to enable those reading the findings of evaluation where the techniques have been applied to.

1. Beneficiary Surveys for Business Support Services[1]

Description of the technique

When is it appropriate to undertake beneficiary surveys related to policy interventions in order to arrive at some measure of overall impact and specific benefits to the individual firm or groups of firms? These surveys can be an expensive option for policy-makers and it is crucial that they are undertaken with a clear understanding of their strengths and weaknesses.

The main imperative which drives the need for beneficiary surveys is information on the performance of the programme from the simple need to ascertain that participant needs are being met through to metrics around impact and benefit.

Circumstances in which it is applied

The starting point of the design of any beneficiary survey is to understand the rationale for intervention and to fully appreciate the customer experience / journey which will have been affected during the delivery of the business support services. Mapping the customer journey effectively is an important first step in the design of a robust beneficiary survey and should be closely related to the Project Logic Model. This provides clarity on actual inputs and anticipated outputs and outcomes. Another essential feature for a quality beneficiary survey is the existence of a comprehensive CRM system which will provide details of all businesses supported under the intervention. This is important to enable appropriate and robust sample selection – especially if a non-beneficiary survey is required.

Sample Size, Response Rates and Outliers

It is clear that robust guidelines are issued in terms of target sample and associated response rates. A response rate of 70% for a beneficiary survey is normally set as the target and the reported confidence intervals are relatively robust for the main part of the survey – satisfaction rates; estimates of additionality.

One of the main issues that may restrain the usefulness of a beneficiary survey is its representativeness. This is something that can be addressed with careful consideration over achieved sample sizes. A detailed description of the sample profile will also provides a clear indication of how representative the beneficiary survey is with respect to the population of all beneficiaries of a particular Programme.

However, for some of the detailed work on estimated financial benefits from the beneficiary survey there would appear to be some issues with outliers which render the estimates problematic. This is always an issue with self-assessment surveys (especially using CATI)[3]. What is to be done with verified outliers? Apply caution and common sense is the response when deriving aggregate benefits for a particular intervention. Extreme responses, which have been verified, are part of the outcome and should be included in all analysis but they do cause problems[2].

[1] This section of the Sourcebook was written by Professor Mark Hart of Aston Business School, Aston University

[2] This problem arose in the context of the estimated benefits of UKTI trade development support derived from PIMS and many hours were expended discussing the issue. Erring on the side of caution was the outcome.

The main steps involved

Surveys: Structure and Content

What should a beneficiary survey include for Programme evaluation? In brief the following are essential prerequisites with some sample questions:

1. **Awareness and accessibility** of the Programme – i.e., entry into the beneficiary category: *"Thinking about the different ways in which you can contact Programme X, would you say that the service was... [very accessible..........not very accessible]"*

2. **Effectiveness** - satisfaction with the intervention against stated objectives of the programme:

   a) Information received

   b) Workshop content

   c) Quality of business mentors

   d) Advice offered

   e) Referrals to other sources of business support

Some useful questions under these headings include the following:

"Overall, how did your experience of Programme X compare with your expectations? Would you say that your expectations were... [exceeded..........not met at all]?"

"Overall, how satisfied are you with the services you have received from Programme X over the last 12 months......[not very satisfied.........very satisfied]?"

In addition, a series of statements can be included to ascertain how the business found specific aspects of the programme (using a disagree/agree scale):

Do you agree or disagree that . . .
"We received all the support and help that we needed"
"The support we received was not relevant to our business needs"
"We would recommend Programme X to other businesses needing help"
"On balance Programme X had a negative effect on our business"
"Programme X is something we can trust to provide us with impartial advice and support"

"Overall, how satisfied are you with the quality of the advice you have received from the third party organisations and individuals working with Programme X.......[very satisfied.....not very satisfied]?"

[3] Computer-assisted telephone interviewing

3. **Outcomes and Impact** - Here we follow the chain of causality set out in the Programme logic model. In the United Kingdom the approach and methods described are consistent with the guidance set out in HM Treasury's Green Book[3] and the (former) DTI's Impact Evaluation Framework (IEF)[4]. In particular it goes beyond the 'gross' outcomes and impacts generated by a Programme to identify the 'net' effects, considering the counterfactual scenario of what would have happened without the existence of Programme X and taking account of the concept of additionality and its various elements.

   There are two dimensions to this. First, an approach based on the results from a series of self-assessment questions in the survey which asked beneficiaries to indicate the effects (outcomes) of the assistance received from Programme X on their business. The intention here is to use a series of standard questions which would facilitate comparison with previous evaluations and indeed other forms of business support. Second, the development of an econometric model to estimate the effects of Programme X intervention based on a survey of beneficiaries and a non-beneficiary control group.

Self-Assessment Effects

The emphasis in the discussion will be on the following components of that assessment:

a) *Motivations for seeking assistance*

b) *Behavioural effects* – the following table indicates the types of outcomes that can be explored (% reporting Yes would be the metric)

More inclined to use external business support for general information and advice
More inclined to use specialist consultancy services
Image of the business has improved
Technical capacity of the business has improved
Financial management skills of the business has improved
Business is better at planning
Business is better equipped to seek external finance
Business has developed a greater capacity to engage in export activity
Business is better able to deal with regulation and compliance issues
Invested more resources (time and money) in training staff
Business has more capability to develop new products or services
Business has improved the quality of its products or services

A follow-up question on each of the areas of business behaviour could be included to ascertain the extent to which, if a respondent replied that they thought there was a

[3] http://www.hm-treasury.gov.uk/Economic_Data_and_Tools/Greenbook/data_greenbook_index.cfm

[4] http://www.berr.gov.uk/files/file21900.pdf

benefit of Programme X assistance, this impact was a direct result of the Programme (on a scale of 1 not very likely to 5 to a critical extent).

c) *Additionality - d*espite the obvious problems inherent in asking beneficiary businesses the rather hypothetical 'counter-factual' question what would have happened in the absence of assistance this approach has become a consistent feature of the evaluations of business support programmes. There are intrinsic difficulties associated with this technique when used in this regard which is commonly referred to as 'respondents effect', that is, the fact that respondents (firms) may purposely exaggerate (in either an upwards or downwards direction) the impact of assistance from an external influence, such as Programme X. More precisely, respondents may exaggerate the impact of assistance for fear that they may reduce their chances of receiving repeat assistance (if they were not deemed by the development agency as really meriting assistance the first time round). On the other hand, other beneficiaries may be likely to play down the impact of assistance attributing success to themselves and their own personal characteristics (such as own motivation; education; business idea etc). These self-reported additionality questions are set out in the table below.

We would have achieved similar business outcomes anyway
We would have achieved similar business outcomes, but not as quickly
We would have achieved some but not all of the business outcomes
We probably would not have achieved similar business outcomes
We definitely would not have achieved similar business outcomes
(None of these)

d) *Timing of Effects* - a significant proportion of firms, however, anticipate future benefits from Programme X support. This will have clear implications for the interpretation of the results from the standard self-reported additionality questions set out in the previous table – i.e., there will be a tendency towards an underestimation of the overall effects of assistance on the business. This raises the very obvious question of when evaluations should be undertaken.

You have already realised all the benefits
You expect to realise all the benefits in the next year
You expect to realise them in the next 2 years
In the next 3 years
In the next 4 years
In the next 5 years
Or will it take more than 5 years to fully realise all the benefits
(No benefits experienced)

In general, self-reported findings from the beneficiary survey point towards a short-term assessment, that the full benefits of Programme X assistance are restricted to a minority of businesses in the sample. Of importance is to recognise that not all the benefits of Programme X support will have been realised at the time of the evaluation.

Beneficiary Surveys – adding value

While not crucial in obtaining headline impacts and benefits, the absence of a non-beneficiary control group does limit the ability to draw inferences about additionality of the intervention based solely on the self-assessment of the beneficiary. For example, the use of unsuccessful applicants to the Programme as a comparison group offered the opportunity to gather information on two aspects of support:

1) The outcome in terms of performance.

2) Possible alternative sources of support for the project.

Normally, non-beneficiaries (control groups) serve to provide an additional source of information on the assessment of the counterfactual. First, they serve to provide a 'benchmark' for the programme beneficiaries in terms of what would have happened in terms of performance – e.g., employment, sales, exports, R&D expenditure. Second, they are able to assess the extent to which alternative sources of external support (if any) are available for projects which the programme had been designed to support. Control groups are also a core component of any evaluation study which meets IEF guidelines in the United Kingdom[5]. The question to address here is the extent to which they add value to the simple focus on beneficiary surveys.

Related to the issues about controls is the issue of selection. Before we can begin to talk in terms of whether a particular product or service has had a particular benefit for participating firms there is a need to address the issue of selection. Put simply, we need to reach a view on whether the product/service has, for example, high levels of additionality due to better performing firms coming forward for assistance or whether better performing firms are selected into the programme in the first instance. Obviously, a methodology which relies upon beneficiary surveys alone does not satisfactorily address this issue.

Two groups of firms are needed:

Non-Beneficiaries – those which received no Programme X support. In fact, these firms may have received support from the programme in the past – prior support outside the time period for the current evaluation so care needs to be taken over possible contamination effects related to the timing of effects of previous assistance (see above).

Beneficiaries – those which received some support from Programme X. These firms may also have received Programme X support in previous periods that fall outside the period of interest in the evaluation. A well-constructed CRM system can help resolve these contamination effects for both beneficiary and non-beneficiary groups – questions need to be inserted for both groups to ascertain all other forms of business support to help focus on the effects of this specific Programme assistance.

[5] See, for example, BIS (2009*) RDA Evaluation: Practical Guidance on Implementing the Framework*, BIS, December 2009.

To complement/challenge the findings from the self-reported outcomes from a simple beneficiary survey an econometric analysis can be developed to assess whether firms which received Programme X assistance have subsequently performed better than they would have without assistance. This approach obviously requires a non-beneficiary survey which in itself raises another set of challenges as we seek to develop the counterfactual which is NOT reliant on a self-assessment methodology.

The essential question is to determine the effect that Programme X support has on firm performance. In other words the task is to determine whether beneficiaries grow faster than non-beneficiaries as a result of the assistance received. Two main issues arise in estimating the impact of assistance on an individual firm. First, the characteristics of beneficiaries and non-assisted firms may differ substantially suggesting that unless these differences are controlled for in the estimation then any assessment of the effect of assistance is likely to be misleading. This emphasises the importance of a strongly multivariate (econometric) approach which explicitly allows for differences in the characteristics of assisted and non-assisted companies, their strategic orientations and the characteristics of their owner-managers and managerial teams.

Second, previous studies have also emphasised the importance of clearly identifying any selection effect to avoid any potential bias due to the selection by Programme Managers of either better or worse than average firms to assist. For example, beneficiary firms may tend to have more rapid growth in the year before assistance. If this was used as a criterion for selection for assistance this might impart a bias to the econometric results.

Addressing this point is relatively straightforward, and simply involves the estimation of two related statistical models – a model for the probability that a firm will receive assistance and a second model relating to the effect of selection and assistance on business growth or performance. This two step approach allows a clear identification of the 'selection' and 'assistance' effects as well as explicitly allowing for differences between the characteristics of beneficiary and non-beneficiary firms[6].

Beneficiary Surveys – estimating financial benefit

The NatCen report for BIS reports that the department has access to the best practice for self-assessment surveys (NatCen, 2009)[7]. In particular the report concentrated on a thorough investigation and test of the way in which BIS have been asking the 'benefits' question. This has been incorporated into the PIMS survey undertaken by United Kingdom Trade and Investment (UKTI) which is used as an example of good practice and is set out in detail in Annex A.

The beneficiary survey can carry a line of questioning which allows us to derive an estimate of benefit and the UKTI Performance & Impact Monitoring Surveys (PIMS) survey provides two useful examples:

[6] Two recent examples include Hart, M; Driffield, NL; Roper, S and Mole, K (2008) *Evaluation of Regional Selective Assistance (RSA) and its successor, Selective Finance for Investment in England (SFIE)*, BERR Occasional Paper No. 2; and Driffield, N; Du, J; Hart, M; Love, J and Tapinos, S (2010) *A Comparative Evaluation of the Impact of UK Trade & Investment's R&D Programme and Other UKTI Support that Impacts R&D,* UKTI Report, March 2010.et al (2010)

[7] NatCen (2009) Self assessment as a tool to measure the economic impact of BERR policies - a best practice guide. Department for Business, Innovation and Skills (BIS: London)

Considering now JUST the anticipated financial gains to YOUR BUSINESS of the activities of the <Programme X participation/assistance> and in terms of bottom-line profits, would you say that the gains TO YOUR BUSINESS are expected to be greater than the costs, about the same as the costs or less than the costs? PROBE AS PER PRE-CODES

Greater than the costs………………………………………………	1
About the same as the costs………………………………………	2
Less than the costs…………………………………………………	3
(None apply) ………………………………………………………	4
(Don't know) ………………………………………………………	5

I would now like you to consider any financial gain to YOUR BUSINESS, in terms of its bottom line profit, of your < Programme X participation/assistance>. **Please could you estimate what the financial gain will be in £. Please include benefits you expect to experience in the future, even if they've not yet been realised.**

ADD IF NECESSARY: **Please just give me your best estimate. £________________**

Thinking again more broadly about the overall aims and objectives of the < Programme X participation/assistance **>, if the group is ultimately successful in achieving these aims, what impact would you envisage this having on YOUR BUSINESS in terms of bottom-line profits? Would you say…?**

Zero/nothing…………………………………	1
Up to £500……………………………………	2
£501 -£1,000 ………………………………..	3
£1,001 -£2,000……………………………..	4
£2,001 -£5,000……………………………..	5
£5,001 -£10,000…………………………….	6
£10 001 -£20 000 ………………………….	7
£20 001 -£50 000………………………….	8
£50 001 -£100,000 ……………………….	9
£100,001 -£500,000…………………….	10
£500,001 -£1million …………………….	11
£1million -£5million ……………………….	12
More than £5million …………………….	13
(Don't know)…………………………………	14
(Refused)……………………………………….	15

Can I just check, is that figure of £ xxx the expected gain in terms of the increase to your bottom-line profit, or the increase to your turnover?

Bottom line profit……………………….	1
Turnover……………………………………..	2
(Don't know)………………………………	3

However, having sat through many pilot interviews (using CATI) for UKTI we need to proceed with extreme caution. The caveats of the approach are well rehearsed in the NatCen report and within the constraints of the self-assessment approach the suggested questions will add significantly to our understanding of the measures of benefit that will result.

The problem here is what we obtain from the respondent even through these revised questions using a CATI methodology. Are the answers any more than an educated guess? Face-to-face interviews are crucial here to develop an understanding of benefit and the cognitive processes that led to that self-assessment. Certainly, the PIMS approach of going back to respondents and presenting them with their previous answer to the financial benefit question and seeking verification is an important option to adopt.

One should not rely too heavily on this as the sole source of evidence on the 'hard' measures of benefit and to construct complementary ways in which these estimates can be derived. I set these out in the concluding section.

Strengths and limitations of the technique

Beneficiary surveys have been used for many years to ascertain information from programme beneficiaries about their satisfaction and the overall benefits and impact of the intervention. They provide useful information to policymakers in the context of the rationale for intervention as represented in the Logic Model but there are many ways they can be enhanced to demonstrate their added usefulness in an environment when there are fewer resources available for business support.

Beneficiary Surveys - Issues

Beneficiary surveys can be improved in order to serve as a more robust internal source of 'early-stage' impact and benefit measures. This can be done by the following:

Ensure there is a non-beneficiary component – or at least for every time a new product/service is introduced into the survey. If focused on unsuccessful applicants this can provide important information on how the identified projects have been taken forward by businesses and, in particular, the alternative forms of support that have been accessed.

Develop a longitudinal dimension to the Beneficiary surveys to allow for multivariate analysis using pooled cross-sectional data. Sample size is important for regression analysis and this would ensure that there are sufficient respondents in the dataset for such an approach.

To complement the value of beneficiary surveys it may be worth investigating the role of linking beneficiaries and non-beneficiaries to ONS administrative and commercial datasets to obtain performance data for monitoring and evaluation purposes. The financial information required for the benefit calculations can be obtained more efficiently by this method and would be more robust than the reliance on self-reported responses from owners/senior managers.

However, the inclusion of a non-beneficiary survey is perhaps not the most efficient way to establish a counterfactual. There is the obvious issue of cost as well as the fact that performance data can be contained by other means – for example, the development of a data-linking approach to monitoring business support products and services would provide performance outcomes for beneficiaries, unsuccessful applicants as well as a range of other comparisons groups if required.

Over-Reliance on Descriptive Analysis

The analysis of many beneficiary surveys is mainly descriptive in nature ranging from the profile of the respondents to the self-assessment of additionality and the ultimate measure of impact and benefit. This limits the interpretation of the results as each response from beneficiaries is viewed in isolation. When reviewing the analysis there is a constant question in the background about the overall impact assessment for the particular product/service and how this interacts with other variables such as size and sector simultaneously.

Multivariate analysis can be effectively introduced into the reporting framework for beneficiary surveys. This would allow a much more detailed assessment of the effects of participation/treatment – what types of firms benefit the most.

Indeed, when the UKTI Performance & Impact Monitoring Surveys (PIMS) quarterly data was made available for this type of analysis the immediate contribution to the discussion of the way trade development activities produced effects on R&D expenditure was quite powerful[8]. With over 4,000 observations from consecutive waves of PIMS the probit regressions (dependent variable being one of UKTI's measures – "Increased R&D Expenditure" it was possible to see the effects of size, sector and region as well as other key variables such as innovation and length of time exporting overseas[9].

Further, the secondary analysis of the PIMS dataset was able also to consider the relative importance of particular interventions (Passport to Export compared to TAP for example) while controlling for a range of firm characteristics. Indeed, in the formal evaluation component of the project for UKTI to evaluate the effects of trade development activity on R&D expenditure using a beneficiary/non-beneficiary survey which was linked to the PIMS data, we were able to develop a narrative on the intensity of support received from UKTI and the R&D outcome. For example, the maximum impact of trade development on the R&D outcome was positively associated with between 5-9 separate interventions by UKTI.

United Kingdom: UKTI - PIMS – A Case Study

Introduction: What is PIMS?

UKTI's Performance & Impact Monitoring Surveys (PIMS). PIMS has been developed in response to the need for consistent monitoring data across all the key UKTI trade development products and services, which will in turn enable UKTI to improve its own performance and the value for

[8] Indeed, there was a UKTI ITT issued in July 2010 which sought secondary regression analysis of PIMS in order to determine the links between the public and private benefits of exporting, and certain firm characteristics.

[9] See Driffield, N; Du, J; Hart, M; Love, J and Tapinos, S (2010) *A Comparative Evaluation of the Impact of UK Trade & Investment's R&D Programme and Other UKTI Support that Impacts R&D,* UKTI Report, March 2010.

money it provides. The primary research objectives are to provide evidence of the impact and effectiveness of the various trade development programmes, and provide data for a number of key survey-based measures.

The main PIMS programme (commencing with PIMS 1 in 2006) takes the form of quarterly surveys of businesses that have recently received support from UKTI, and focuses on the *anticipated* impacts and outcomes of this support (i.e. the benefits that firms expect to see in the future as a result of the assistance they have received). However, PIMS also includes a smaller-scale follow-up stage, where a sample of those firms interviewed in the main PIMS wave are contacted again approximately 10 months after the initial survey to further explore the impact of UKTI's support. The primary purpose of these follow-up surveys is to explore the *actual* impacts and outcomes of the support to date (i.e. the benefits that firms have already realised as a result of the assistance received).

The research was conducted by an independent research survey company and utilised a telephone methodology, with all interviews conducted using CATI[10]. The final questionnaire averaged 20-25 minutes and a copy of one of the most recent questionnaires is appended to this report.

Whilst the questionnaire included a significant amount of text substitution and routing to ensure that questions were relevant to the specific type of support that each business had received, the core of the questionnaire was kept consistent across all product groups in order that comparable data was collected. Where appropriate, respondents were reminded of the answers they gave previously in PIMS surveys. The questionnaire is fully piloted prior to the start of the fieldwork both quantitatively and qualitatively, and these pilots checked the flow, clarity, relevance and length of the questionnaire as well as the content.

Currently, the PIMS quarterly surveys are in waves 22-24 and a large dataset has now been constructed over the last 8 years. The latest report has been published in October 2011 – but all the PIMS survey results (PIMS 1-24) and the questions are available at the following website:

http://www.ukti.gov.uk/uktihome/aboutukti/ourperformance/performanceimpactandmonitoringsurvey/quarterlysurveys.html

The next example shows how on its own a beneficiary dataset can be used to assist policymakers focus marginal resources.

Analysis of Performance and Impact Monitoring Survey (PIMS) data for UKTI trade services: a summary

The purpose of this analysis was to isolate any differences across services in reported R&D effects, and to identify client characteristics most likely to be associated with reported R&D effects. As this analysis was based solely on evidence from users, captured via PIMS, it did not test the validity of the reported R&D effects.

Use of multivariate analysis, for the first time, allowed service effects to be distinguished from differences in client profile across services, which would also affect the likelihood of reporting additional R&D. The analysis covered PIMS waves 6-9, which involved interviews with around 3,000 firms.

[10] Computer Assisted Telephone Interviewing.

The analysis took advantage of the fact that the PIMS dataset contains comparable data by service on details of client profile and export experience, as well as on a range of reported effects of services on the client's business, including increased sales and increased R&D. The measure of increased R&D used in this analysis is the one used by UKTI to report performance against its SR07 Target relating to increased business R&D. The measure relies solely on judgments made by the respondent about their experience, and does not take account of selection effects.

Key findings of the multivariate analysis are:

- Innovative and growing firms, especially in manufacturing, are more likely to report 'increased R&D'.
- Firms reporting 'increased sales' are more likely to report 'increased R&D'.
- In terms of comparing UKTI service effects, the analysis found that some services are significantly more likely to report increased R&D. These are: EMRS, TAP (group), Passport, and UKTI Website (users of the Business Opportunities Alert Service). Weakly significant effects were also found for: Overseas Posts; International Business Specialists, Market Visit Support, and Outward missions.

These findings broadly correspond to the pattern of impact which appears in the published bivariate PIMS results. However, two services – ECR, and advice provided by teams in the English regions to 'New to Export' clients – show comparatively less likelihood of reporting additional R&D when client profile is controlled for through the multivariate analysis. The absence of a non-beneficiary survey of similar size prohibits a more robust econometric analysis but nevertheless the results are more informative than simple descriptives.

2. Case studies

For the reader interested in more detail on qualitative case studies and the various approaches to them, see Vanclay (2011), also on the INFOREGIO website, at: http://ec.europa.eu/regional_policy/sources/docgener/evaluation/doc/performance/Vanclay.pdf

For the reader interested in carrying out case studies of projects, see DG REGIO's guidance on drafting case studies at: http://ec.europa.eu/regional_policy/archive/cooperation/interregional/ecochange/doc/evaluation_brochure_062008_en.pdf

Description of the technique

The case study method involves in-depth study of a phenomenon in a natural setting, drawing on a multitude of perspectives. These multiple perspectives may come from multiple data collection methods (both qualitative and quantitative), or derive from multiple accounts of different actors in the setting. The phenomena may concern individuals, programmes, organisations, projects, groups of people or decision-making processes. Case studies are described as embedded where there is more than a single focus or unit of analysis.

Case studies are information rich. They build up very detailed in-depth understanding of complex real-life interactions and processes. The defining feature of the case study is that it is holistic, paying special attention to context and setting. The case study may be a single case, or it may include multiple cases. Provided resources are adequate, multi-site case studies provide rich opportunities for theoretically informed qualitative evaluation.

Case studies raise a number of issues at the design stage. What will count as a 'case'? What is the basis for selecting cases, and how many? What units of analysis will be included within the case, and how must the data be organised to allow meaningful comparisons to be made? What kind of generalisation is possible?

The purpose of the technique

Case studies are used for the following:

- illustration: the case study is a tool that may be used to add realism to an evaluation if it is presented in a narrative form. The case must, however, be chosen carefully because it must be representative of the programme as a whole or illustrate a specific point - for example a particularly effective action or an approach which was found to have serious deficiencies and which should therefore be avoided in future.
- exploration: putting forward hypotheses for future investigations, identifying the various points of view of the stakeholders.
- critical analysis: verify and validate a statement concerning a programme, project or strategy.
- analysis of implementation: examine the diffusion of services and its mechanisms, often in different places.
- analysis of the impacts of programmes: understanding the nature of the processes producing impacts.

The results of a case study are always presented in a narrative form, as a story, thus giving the reader an "inside view" of the case studied and an impression of authenticity. The case study therefore has an analytical and communicative aim. Readers are more likely to relate to cases where the programme(s) and personnel involved are identified. However, particularly when a case study is being used to illustrate and learn from failures it may be necessary to anonymise some or all of the material in order to secure access to data and personnel.

Case studies can often be designed in a cumulative way to help to answer evaluation questions. The same case programme may also be studied over time to provide an analysis that is updated iteratively. Cases may be descriptive, normative or designed to show causality. They may be particularly useful in pedagogic/training situations for example being used to enable officials to evaluate alternative evaluation methodologies.

Circumstances in which it is applied

Case studies are used extensively in evaluation. Today this method is known to provide valid information for both the evaluation of programmes and the diffusion of new knowledge. Case studies which use sophisticated selection procedures (e.g. "multiple case studies with replication design") tend to replace large-scale quantitative surveys carried out in diverse cultural contexts.

The case study is a method of holistic analysis applied to complex situations. This means that its use is appropriate for the in-depth understanding of behaviours and social phenomena, by using the persons and organisations analysed as a frame of reference. Case studies are valuable for identifying the effects of programmes inductively, by developing assumptions concerning the phenomena linking cause and effect. These assumptions must then be supported by information drawn from the different case studies and testing through the search for alternative explanations.

This may prove useful for observing expected results, but also for revealing unexpected ones. The method is less suited to the identification of causal links, although it may be used to demonstrate that they are likely to exist.

The case study is intended to be the most complete illustration possible of a given situation, so as to give a precise image of current phenomena and to understand their causes. This is obtained by means of the description and then the analysis of examples situated in their context. It follows that this type of analysis must be based on multiple data sources, such as interviews, observations over time, statistics, physical information, etc. The data must also be cross-checked in order to ensure its coherence. The notion of "context" encompasses all the factors that could affect the case studied. Thus, for example, the impacts of a specific project on the beneficiaries are influenced by a large number of external factors.

The multiple case study method is particularly well suited to analyses of the various member States and regions, but also to thematic evaluations. The flexibility of each case study makes it possible to draw up an adequate description of the peculiarities of a given place or a project. The formulation of a common set of questions, relative to the evaluation, facilitates the analysis of the results obtained from multiple case studies. In fact the results prove to be more sound when they are produced in relation to a variety of places (through re-using case studies). Similarly, the specificity of success stories or failures will then seems more obvious. It should be recalled that the transversal analysis of cases consists of cross-referenced qualitative examinations and a description based on the frame of reference established by the evaluation questions. A cumulative process may be sought when the evaluation is focused, for example, on operational programmes in less developed regions and when the conclusions have to be synthesised on the scale of several member States.

The presentation of the results of several case studies could be a barrier to more generalised use. This difficulty may be solved by means of a graphic summary providing a brief report of the case history and a graphic presentation of the results, in relation to each of the questions. In this form, the answer to each of the evaluation questions, for each case, is set out on a single page: a graphic presentation at the top, a short but rich summary of the case history, with the main results, and a concise conclusion. The summary of the transversal case could be sketched in the same way, followed by conclusions and recommendations.

Another type of case study that may be applied to the Structural Funds is the "integrated approach". This approach takes into account, for example, the study of results in the context of a specific programme.

The main steps involved

The quantity of work required by a case study may vary widely. One must bear in mind that the case study must be sufficiently rich to give the reader an impression of what actually occurred. However, the case study is part of the least standardised methods and may encompass a range of different methodologies in different situations. Carrying out a case study involves the following steps:

Step 1. Selection of cases to study

There are at least three criteria for selecting cases: convenience/access, the purpose to which they are to put and the extent to which they can be considered to provide wider insights beyond the particular case in question. The selection of cases is a critical step for generalising and answering evaluation questions. It is difficult to justify a selection based only on convenience (easy access to data) and probabilistic surveys are sometimes difficult to carry out. Thus, a choice based on purpose is appropriate in most cases.

Step 2. Data collection and process

Theoretically, data collection covers all available information about a case including that derived from project documents, project meeting reports, and collected at the various operational levels: interviews with project leaders and staff; observation of the site of the project; surveys among the beneficiaries of the services provided by the project. These data must be collected, recorded (compilation of a "register") cleaned and pieced together so that they can be used in the final report.

Step 3. Case report

Drawing up the report on the case involves the organisation of all the raw data on the case into a body of exploitable information. This is then edited, redundant information is eliminated, and the different parts are combined. The report is organised in such a way as to be easy to consult, either chronologically or thematically. The report must include all the information required for subsequent analysis, that is to say, for constructing an account of the case study.

Step 4. Account

The case monograph should give the reader immediate access to relevant information and to the particular situation of the case - the situation of a project - and provide an understanding of the project as a whole. Each case study, in an evaluation report, must be isolated (the size may vary between one and five pages). Nevertheless, in the last steps of the analysis the cases may be used as contrasts or comparisons, depending on the evaluation objectives.

Strengths and limitations of the technique

- The case study is relevant for giving a view of processes and complexities that are impossible to see in any other way. It may even make outside persons, such as European managers who are hardly involved in this field, aware of the reality of daily actions. It provides them with a clearer view of the way in which the programme is put into practice once the decision has been taken with the national authorities.

- Case studies may permit a different kind of generalisation than one based on probabilistic sampling and tests of statistical significance. Case study designs that balance depth and breadth, and are purposefully sampled, will allow the evaluator to make extrapolations, or modest speculations, about the likely applicability of findings to other situations under similar, but not identical, conditions. Sampling strategies should be planned with the stakeholders' desire for extrapolation in mind.

- This approach is less appropriate for measuring the amplitude of impacts or for inferring the causality.

- Due to the cost of setting up a good case study (requiring sources of multiple data and competent evaluators), it is necessary to limit the number of observations. The case study may, however, be re-used and applied to other context, thus providing economies of scale.

- The credibility of the results of the case study is likely to be undermined if the method is not implemented correctly: incompleteness, arbitrary selection of information, comments cut short, distortion of results, etc. To enhance the reliability of a case study, several precautions are recommended, e.g.: rereading of the case studies by the persons concerned in order to verify the precision and the veracity of the data and their interpretation; or having two different evaluators write down their comments on the same case; involving outside professionals (such as journalists) in the writing of the comments.

-

Bibliography

Dufour S., Fortin D., Hamel J. (1994), 'L'enquête de terrain en sciences sociales: l'approche monographique et les méthodes qualitatives', Montréal: Saint-Martin

GAO (1990), 'Case Study Evaluations.' Washington DC: General Accounting Office, pp133 - Sound general approach to the method.

Shaw, I. (1999), Qualitative Evaluation. London: Sage Publications.

Technopolis, (2008): Analysing ERDF co-financed innovative projects – Case Study Manual

Vanclay F (2012): Guidance for the design of qualitative case study evaluation, available at: http://ec.europa.eu/regional_policy/sources/docgener/evaluation/doc/performance/Vanclay.pdf

Yin R.K. (1994), 'Case Study Research - Design and Methods', 2nd ed. Newbury Park, Sage Publications – a Complete guide to the case study technique.

3. Cost benefit analysis

Description of the technique

Cost-benefit analysis (CBA) is a method of evaluating the net economic impact of a public project. Projects typically involve public investments, but in principle the same method is applicable to a variety of interventions, for example, subsidies for private projects, reforms in regulation, new tax rates. The aim of CBA is to determine whether a project is desirable from the point of view of social welfare, by means of the algebraic sum of the time-discounted economic costs and benefits of the project.

The technique used is based on:

1. forecasting the economic effects of a project.

2. quantifying them by means of appropriate measuring procedures.

3. monetising them, wherever possible, using conventional techniques for monetising the economic results.

4. calculating the economic return, using a concise indicator that allows an opinion to be formulated regarding the performance of the project.

The purpose of the technique

The justification for an investment project tallies with the feasibility and economic performance.

Cost-benefit analysis usually accompanies a feasibility study (technical, financial, legislative, organisational) of the project itself and it constitutes the final synthesis.

The main advantage of CBA compared to other traditional accounting evaluation techniques is that externalities and observed price distortions are also considered. In this way market imperfections are explicitly considered, which are reflected neither in corporate accounting nor, as a rule, in national accounting systems.

Circumstances in which it is applied

The first ideas and applications of CBA can be traced back to nineteenth century France, and later they spread to the UK and USA, especially in the transport and hydraulic works sectors. The systematic use of cost-benefit analysis was developed by international organisations, especially the World Bank. Today cost-benefit analysis plays an important role in evaluating major infrastructure projects, especially those that are co-financed by the ERDF and the Cohesion Fund, and it constitutes a requisite for European Community co-financing.

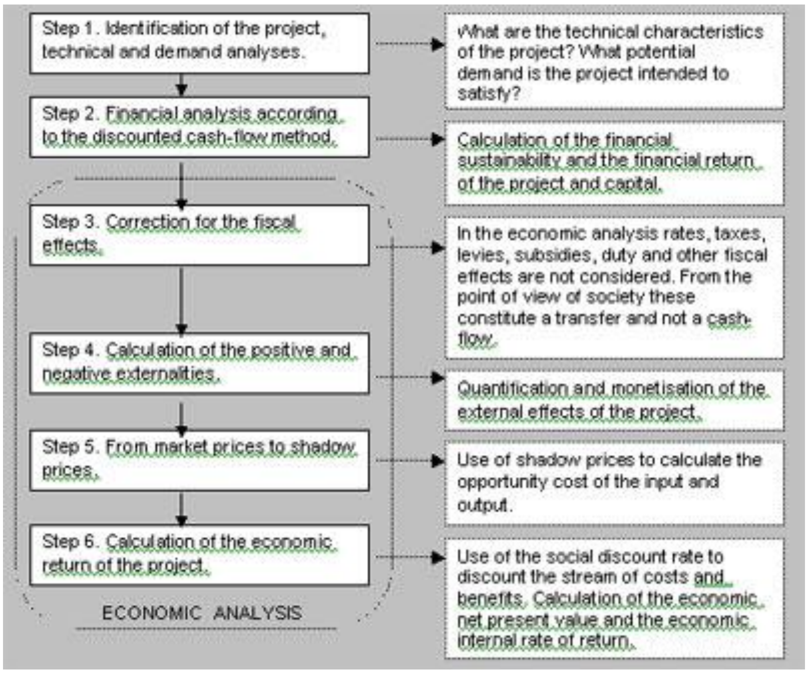
Generally speaking cost-benefit analysis is used in the ex-ante evaluation for the selection of an investment project. It can also be used ex-post to measure the economic impact of an intervention. It is used when the effects of an intervention go beyond the simple financial effects for the private investor. It is normally used for major infrastructure projects, especially in the transport and environment sectors, where it is easier to quantify and monetise the non-market effects. CBA is also used to evaluate projects in the health, education and cultural heritage sectors.

CBA is not normally used to evaluate programmes and policies, even though in principle it could be used to study the effect of changes in specific political parameters (for example customs tariffs, pollution thresholds, etc.).

The main steps involved

Basically cost-benefit analysis is made up of three parts:

- a technical-engineering part in which the context and technical characteristics of the project are identified;
- a financial analysis that represents the starting point for the CBA and that leads the analysis from the point of view of the private investor;
- an economic analysis, the true core of CBA, which, starting with the financial analysis that serves to identify all the income and expenditure items and the relative market prices, applies a series of corrections that allow us to pass from the point of view of the private investor to that of the public operator.



Step 1. Identification of the project, technical and demand analyses.

What are the technical characteristics of the project? What potential demand is the project intended to satisfy?

Step 2. Financial analysis according to the discounted cash-flow method.

Calculation of the financial sustainability and the financial return of the project and capital.

Step 3. Correction for the fiscal effects.

In the economic analysis rates, taxes, levies, subsidies, duty and other fiscal effects are not considered. From the point of view of society these constitute a transfer and not a cash-flow.

Step 4. Calculation of the positive and negative externalities.

Quantification and monetisation of the external effects of the project.

Step 5. From market prices to shadow prices.

Use of shadow prices to calculate the opportunity cost of the input and output.

Step 6. Calculation of the economic return of the project.

Use of the social discount rate to discount the stream of costs and benefits. Calculation of the economic net present value and the economic internal rate of return.

Step 1. Identification of the project, technical and demand analyses.

The first step serves to place the project in its implemental context.

Obviously it is necessary to identify the object of the evaluation, the unit of analysis to which the cost-benefit analysis is applied. This is particularly important for groups of projects or for parts or phases of a larger project that have their own planning autonomy.

Identifying a project also means clearly defining the socio-economic objectives that the project intends to achieve.

The technical analysis is carried out to ensure the feasibility of the projected work from a technical point of view. This involves aspects of an engineering, management, localization, marketing and organisational nature. The proposed project must show that it is the best of the possible alternatives.

For each project at least three alternatives may be considered:

- the do nothing alternative
- the do minimum alternative
- the do something alternative.

Step 2. Financial analysis

The financial analysis is the starting point for the subsequent economic analysis. It provides all the necessary data regarding input, output, their relative prices and how they are distributed over time. It serves to:

- formulate the tables for the analysis of the cash-flows (selection of the important cost and revenue items)
- evaluate the financial feasibility (verification of sustainability)
- evaluate the financial benefit by calculating the return from the private investor's (financial return of the project and the capital) point of view.

Financial feasibility is an essential condition for the viability of the project, but financial convenience is not necessary: On the contrary, if a project is extremely convenient for a private investor then there is less need to decide on the convenience of public financing. The financial analysis is carried out using the discounted cash flow method. The choice of the discount rate is crucial to the assessment of the costs weighted against the benefits over a longer period of time. The discount rate is the rate by which benefits that accrue in some future time period must be adjusted so that they can be compared with values in the present. This method considers only the real monetary income and expenditure of the project and does not include accounting conventions such as depreciation, reserves and so on.

The monetary income and expenditure are recorded at the time they effectively occur. Thus it is necessary to define a time horizon that is coherent with the project's life cycle, and to estimate not only the income and expenditure, but how they are expected to break down over the entire time horizon.

In the last year of the time horizon an appropriate residual value is calculated as a percentage of the investment costs. This represents the potential income flows that the project will still be capable of generating even after the time horizon considered. It can be assimilated into the liquidation value of the project.

The financial analysis is made up of three tables that summarize the basic data and three tables for the calculation of important indicators. The initial three tables are:

- investment costs and residual value; this includes the value of the fixed assets, (land, buildings, extraordinary maintenance), pre-production expenses (licences, patents, etc.), variations in working capital (cash, clients, stocks, current liabilities) and residual value, which appears as a single positive item in the last year of the time horizon.
- operating costs and revenue; this includes all the operating costs (raw materials, labour, electricity, maintenance) and any possible revenue items (tariff and non-tariff income);
- sources of financing; this includes private equity, all public contributions (local, national, community level), loans and other sources of financing.

The three summary tables of the financial analysis used for the calculation of the indicators are:

- the financial sustainability table; this includes all the items of the initial three tables. By calculating the balance between revenue and expenditure we obtain a calculation of the accumulated generated cash (the algebraic sum between the balance of the year considered and the cash accumulated up to the previous year. Financial sustainability is ensured if the accumulated generated cash is positive or, at most, equal to zero for all the years considered. On the contrary, if the accumulated generated cash is negative even for just one year, the project is not feasible from the financial point of view and it will be necessary to modify the structure of the project in order to evaluate it.

The table for calculating the return of the project is the composition of the first and second of the initial tables. Expenditure includes all investment and operating costs and revenues include any possible income plus the residual value. By calculating the balances, discounted at an appropriate rate, it is possible to define a financial net present value and a financial internal rate of return.

The table for calculating the return on capital is the composition of the first and third of the initial tables.

Step 3. Correction for the fiscal effects.

In the financial analysis carried out from the point of view of the private investor some items are included, like for example taxes on profits, that represent neither a social benefit not a cost, but rather a transfer from one social group to another. Other examples of fiscal effects can be found in subsidies, in welfare contributions considered in the cost of labour and the effects of duties on the prices of inputs and outputs. In this step two types or corrections are carried out:

- all fiscal items (taxes, subsidies) are eliminated;
- market prices are modified whenever they reflect effects of a fiscal nature, such as duty, VAT and other indirect taxes (this type of correction is assimilable to the one carried out in Step 5).

Step 4. Calculation of the positive and negative externalities.

In evaluating the convenience the public operator also considers the externalities generated by the project. The externalities are social costs or benefits that manifest themselves beyond the realms of the project and influence the welfare of third parties without any monetary compensation. As such they are not captured by market mechanisms and are not monetised, since their effects are transmitted through real variables that influence the welfare of individuals and not through price mechanisms. Such effects, which influence the welfare of the social group involved, must be quantified and then monetised in order to be included in the analysis as a true item of input or output.

The external effects generated by a project may be easy to identify, but are often difficult to quantify. After quantifying in physical terms, a monetary value must be attributed to the quantified benefit; this operation requires a lot of approximations and recourse to some standard practices that have been consolidated at an international level. This, for example, is the case of any calculation of the value of time or of human lives. In these cases one tries to artificially reconstruct the market mechanism to measure the preferences of individuals via the declared preferences method (the willingness to pay method) or the revealed preferences method (the value is approximated by calculating the savings in spending, as in the health field, or the price of equivalent goods and services). Importantly, to provide a broader understanding of the implications of a project or programme, social, environmental and gender impacts, must also be evaluated. To assist in determining these impacts, tools such as environmental impact assessments and gender impact assessments can be conducted.

Step 5. From market prices to shadow prices.

The last correction is made through the calculation of opportune conversion factors which, multiplied by the market price, give the value of the shadow prices. This correction is necessary because the markets are imperfect and market prices don't always reflect the opportunity cost of a good. If prices are distorted they are not a suitable indicator of welfare.

In order to correct the market prices of inputs and outputs the following are used:

- the marginal cost, for non-marketed goods such as the land, local transport services, etc.;
- the border price, for marketed goods;
- the standard conversion factor for minor non-marketed.

As far as salaries are concerned, the two alternative methods of calculating the conversion factor are:

- using a conversion factor lower than one if faced with high unemployment (by reducing the labour costs the net economic value of the project rises compared to the financial analysis);
- calculating an income multiplier that captures the positive external value of creating jobs.

Since the reference price in the economic analysis is the opportunity cost (that is the best alternative use of a specific resource) it is obvious that if faced with high unemployment (as is mainly the case with development projects) the conversion factor will be less than one (the alternative use of the labour resource would be unemployment). If the contrary is true, the conversion factor is greater than one, which means that probably the project diverts labour resources from more productive occupations.

Step 6. Calculation of the economic return of the project.

Having made the three corrections we have built a table of the economic analysis that combines the items contained in the first and second initial tables duly corrected with the elimination of the fiscal effects, and the addition of the external effects and the correction of the prices using discount coefficients. In order to measure the economic convenience, after having time-discounted with a social discount rate (generally different from the financial one), it is now necessary to calculate the net present value and the economic internal rate of return, following the methodology already adopted for the financial analysis.

The economic internal rate of return is expected to be higher than the rate of financial return. If this is not so, then the project would be more convenient for a private investor than for a public operator (unless there are considerable social benefits that are not monetisable).

The calculation of the economic indicators allows for the creation of a ranking of projects and helps in the selection of more than one alternative intervention.

Strengths and limitations of the technique

Strengths of cost benefit analysis

- enables us to express an opinion on the economic-social convenience of a project;
- enables us to create rankings among projects;
- encourages the practice of identifying the economic benefits and costs, even of they are not immediately monetisable.

Limitations of cost benefit analysis

- does not take redistributive effects into consideration (for these one can use a multicriteria analysis);
- does not consider the effect on the economic return of non-monetisable benefits or costs;
- sometimes uses discretional criteria for the monetisation of the costs and benefits for which no market exists.

For all of these reasons cost-benefit analysis is a useful tool for evaluating and selecting projects, but it requires strictness and methodological coherence in its application.

Bibliography

Belli, P., Anderson, J. R., Barnum, H.N, Dixon, J. A., Tan, J-P, 2001, Economic Analysis of Investment Operations. Analytical Tools and Practical Applications, WBI, World Bank, Washington D.C.,

Brent, R.J., 1996, Applied cost-benefit analysis, Cheltenham (UK), Edward Elgar.

CSIL, Milan & DKM, Dublin, 2012, Ten Projects Observed, DG REGIO – an example of ex post cost benefit analysis carried out 5 to 10 years after project completion, available on INFOREGIO at http://ec.europa.eu/regional_policy/sources/docgener/evaluation/pdf/9499_final_report_091112.pdf

Dinwiddy C., Teal F., Principles of cost-benefits analysis for developing countries, Cambridge University Press, 1996.

Economic Development Institute, 1996, The economic evaluation of projects, World Bank, Washington DC.

G. Gauthier, M. Thibault, 1993, L'analyse coûts-avantages, défis et controverses, HECCETAI, Economica.

Evaluation Unit, DG Regional Policy, European Commission, 2008, Guide to cost-benefit analysis of investment projects (Structural Fund-ERDF, Cohesion Fund and IPA). At:

http://ec.europa.eu/regional_policy/sources/docgener/guides/cost/guide2008_en.pdf

Florio, M (ed), 2007, Cost Benefit Analysis and Incentives in Evaluation: The Structural Funds of the European Union,Edward Elgar Publishing Ltd.

Florio, M. and Vignette, S. Cost benefit analysis of infrastructure projects in enlarged EU: an incentive-oriented approach, paper presented ad the Fifth European Conference on Evaluation of the Structural Funds, "Challenges for Evaluation in an Enlarged Europe", Budapest 26/27 June 2003

Jorge, J. and de Rus, G. Cost benefit analysis of Investments in Airport Infrastructure: A Practical Approach, paper presented ad the Fifth European Conference on Evaluation of the Structural Funds, "Challenges for Evaluation in an Enlarged Europe", Budapest 26/27 June 2003

Kirkpatrick, C., Weiss, J., 1996, Cost-benefit Analysis and Project Appraisal in Developing Countries, Elgar, Cheltenham.

Kohli, K.N., 1993, Economic analysis of investment projects: A practical approach, Oxford, Oxford University Press for the Asian Development Bank.

Layard R., Glaister S. (eds), 1994, Cost-benefit Analysis, 2nd edition, Cambridge University Press.

Saerbeck R., 1990, Economic appraisal of projects. Guidelines for a simplified cost-benefit analysis, EIB Paper no.15, European Investment Bank, Luxembourg.

Shofield J.A., 1989, Cost-benefit analysis in urban and regional planning, Allen & Unwin, London.

Glossary

Cost-benefit analysis: a theoretical approach applied to every systematic quantitative evaluation of a public or private project, in order to determine if, and to what extent, the project is convenient from a public or social perspective.

Discounting: Put simply, the discount rate is a percentage used to discount the future value of money. It is used to project your costs into the future, but price them with today's value of money.

Economic analysis: an analysis conducted by using economic values that express the value that society is willing to pay for a good or service. In general the economic analysis assesses goods or services at their use value or their opportunity cost for society (often a border price for tradable goods). It has the same meaning as cost-benefit analysis.

Financial analysis: allows for the accurate forecasting of which resources will cover the expenses. In particular it enables one to: 1. verify and guarantee cash equilibrium (verification of financial sustainability); 2. calculate the indices of the financial return of the investment project based on the net time-discounted cash flows, which refer exclusively to the economic unit that implements the project (firm, managing body).

Traded goods: goods that can be traded internationally in the absence of any restrictive trade policies.

Non-traded goods: goods that cannot be imported or exported, for example local services, unskilled labour and land. In the economic analysis non-marketed goods are assessed at the value of their marginal return if they are intermediate goods or services, or according to the willingness to pay criterion if they are final goods or services.

Socio-economic costs and benefits: opportunity costs or benefits for the economy as a whole. They may differ from private costs to the extent that effective prices differ from shadow prices (social cost = private cost + external cost).

Opportunity cost: the value of a resource in its best alternative use. For the financial analysis the opportunity cost of an acquired input is always its market value. In the economic analysis the opportunity cost of an acquired input is the value of its marginal return in its best alternative use for intermediate goods or services, or its use value (measured by the willingness to pay) for final goods or services.

Willingness to pay: the amount consumers are willing to pay for a good or service. If a consumer's willingness to pay for a good exceeds its price then the consumer enjoys a rent (consumer surplus).

Distortion: condition in which the effective market price of a good differs from the efficient price it would have in the absence of market failures or public policies. This generates a difference between the opportunity cost of a good and its effective price, for example in a monopoly regime, when there are externalities, indirect taxes, duty, tariffs, etc.

Externalities: effects of a project that extend beyond the project itself, and consequently are not included in the financial analysis. In general an externality exists when the production or consumption of a good or service by one economic unit has a direct effect on the welfare of the producers or consumers in another unit without compensation. Externalities may be positive or negative.

Conversion factor: a number that can be multiplied by the national market price or use value of a non-marketed good in order to convert it into a shadow price.

Border price: the unit price of a marketed good at the country's border. For exports this is the FOB (free on board) price and for imports it is the CIF (cost, insurance and freight) price.

*Shadow price: t*he opportunity cost of goods, usually different from the actual market price and from regulated tariffs. It should be used when analysing a project to better reflect the real cost of the inputs and real benefits of the outputs for the society. Often it is used as a synonym of accounting prices.

Economic Rate of Return (ERR): index of the socio-economic profitability of a project. It may differ from the financial rate of return (FRR) due to price distortions. The economic rate of return implies the use of shadow prices and the calculation of a discount rate at which the benefits of the project equal the present costs, that is the economic net present value is equal to zero.

Internal rate of return: the discount rate at which a stream of costs and benefits has a net present value of zero. We speak of financial internal rate of return (FIRR) when the values are estimated at current prices, and economic rate of return (EIRR) when the values are estimated at shadow prices. The internal rate of return is like a reference value to evaluate the results of the proposed project.

Discount rate: the rate at which future values are discounted. The financial and economic discount rates may differ, in the same way in which market prices may differ from shadow prices.

Net Present Value (NPV): the discounted monetary value of the expected net benefits of the project. The economic net present value (ENPV) is different from the financial net present value (FNPV). It is the measure that is often used to determine whether a programme / project is justifiable on economic principals. To calculate NPV, monetary values are assigned to benefits and costs, discounting future benefits and costs using an appropriate discount rate and subtracting the sum total of the discounted costs from the sum total of the discounted benefits. NPV is based on the principle that benefits accruing in the future are worth less than the same level of benefits that accrue now. Furthermore, it takes that view that costs occuring now are more burdonsem that costs that occur in the future. If the NPV is positive, then the financial return on the project is economically acceptable. If the NPV is negative, then the project is not acceptable in purely economic terms.

Residual value: the net present value of the assets and liabilities in the last year of the period chosen for evaluation.

Do nothing / Do minimum / Do something alternatives: If used ex-ante, a cost benefit analysis of a project or intervention, can enable policy makers to assess the feasibility of the projected work from a technical point of view. As a result of this assessment, policy makers should be able to determine whether the intervention is required. The three scenarios above will be a result of the ex-ante cost benefit analysis. Thus, a decision will be made to either do nothing (no intervention / project), intervene in the least possible way, or proceed with the proposed intervention / project. The do nothing option is rarely the solution.

4. Cost effectiveness analysis

Description of the technique

Cost-effectiveness analysis (CEA) is a method that can help to ensure efficient use of investment resources in sectors where benefits are difficult to value. It is a tool for the selection of alternative projects with the same objectives (quantified in physical terms). EA can identify the alternative that, for a given output level, minimises the actual value of costs, or, alternatively, for a given cost, maximises the output level. For example, the evaluator can compare by simple output/cost ratios different projects that aim to lower the fertility rate, or different methods of teaching reading and writing, or different interventions to lower the infant mortality rate.

CEA is used when measurement of benefits in monetary terms is impossible, or the information required is difficult to determine or in any other case when any attempt to make a precise monetary measurement of benefits would be tricky or open to considerable dispute. It does not consider subjective judgments and is not helpful in the case of projects with multiple objectives. In the case of multiple objectives a more sophisticated version of the tool could be used, the weighted cost-effectiveness analysis, which gives weights to objectives to measure their priority scale. Another alternative is a multicriteria analysis. The technique, which looks at the cost of an intervention, and relates it the benefits created, is also closely related to the use of a Value for Money Assessment. Notably, when assessing the value of an intervention, value for money does not necessarily mean achieving outcomes at the lowest cost.

The purpose of the technique

The objective of CEA is to evaluate the effectiveness of a project, that is, its capacity to achieve the desired objectives. The latter should be defined in physical and not monetary terms, e.g. reduction in the morbidity rate through an intervention in the health sector, in relation to the costs incurred to reach them. CEA is best used to decide which alternative maximises the benefits (expressed in physical terms) for the same costs or, vice versa, which one minimises costs for the same objective. The cost-effectiveness ratio allows projects to be compared and ranked according to the costs necessary to achieve the established objectives. Since the objectives cannot be converted into a common numeraire or accounting unit, CEA cannot be used to decide on a project taken in isolation, nor to decide which of two projects would give the better return in two different contexts.

CEA is also used as an alternative to a cost-benefit analysis when social benefits and costs are difficult to monetise, but with strong limitations.

Indeed while a programme may be highly effective at meeting its objectives, it may not to provide good value for money. For example, the programme may be relatively inefficient and the objectives could have been met using fewer resources if an alternative method had been adopted. Assessing the cost effectiveness of a project or programme will not by itself, even when benchmarked against ratios derived from comparable programmes, provide a clear assessment of its social net benefits. It should include additional work to assess the different perceptions of success of different interest groups, as well as an assessment of economy. The value for money approach will often be informed by both top-down and bottom-up analyses. It enables the multiple objectives of a regeneration programme, for example, to be explicitly included, as well as an assessment of the efficient use of resources.

Circumstances in which it is applied

CEA is mainly a tool for the selection of projects within a well defined programme. It has been most commonly used in the evaluation of projects in the health sector, and in all cases where the benefits produced (or the objectives achieved) are difficult to monetise. It is used to make comparisons between alternatives that have the same scope. It cannot be used for projects with different objectives or for a project with multiple objectives. CEA can be used in both ex-ante and ex-post evaluation. It is used when all the expected effects have been defined and are homogeneous and/or can be measured in terms of a key result (e.g. the number of jobs created, the number of trainees, the number of prevented infections). It is used often for evaluating social projects (education, health).

Value for money assessments can be conducted both ex-ante and ex-post. It is used when all the expected effects have been defined and can be measured in terms of a key result (e.g. the number of jobs created, number of business start-ups/survivals). For example, if a Government department were considering the implementation of an initiative that provides assistance to small businesses, they could conduct an ex ante evaluation to identify the likely benefits and the cost per benefit of the scheme, prior to implementing it. The findings of such an evaluation would inform the decision to implement or scrap such a policy. Ex ante, provides the evaluator with the necessary information to assess the cost of the inputs, and the outputs and outcomes of a given policy. This enables them to make a judgment on whether the said scheme provided good returns for the amount of money inputted. Value for Money assessments are common in the UK, closely linked to the UK government's Modernising Government agenda which focuses on the achievement of Best Value for service provision and procurement. Best value is an approach that aims to ensure that public bodies / delivery agents play their part in delivering high quality provisions in a cost-effective manner (both internally and when procuring work).

Example: Cost effectiveness analysis for the choice of technology

The choice of technology is a common example of a situation when benefits can only be measured in some nonmonetary terms. This is the case of determining the minimum cost for a given output.

Case study: Improvement of boilers in a district heating system.

Three technological alternatives are shown below:

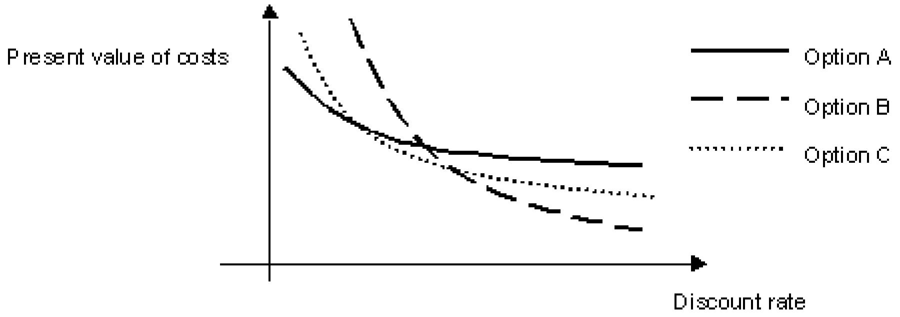
- Technology A: replacement of all existing boilers with new woodburning ones
- Technology B: renovation of existing oil-and gas-fired boilers
- Technology C: essential repairs of existing boilers

Total costs for each projects are:

Total costs Thousand Euro	Investment cost	Annual fixed maintenance costs	fuel cost
Technology A	6000	150	300
Technology B	3250	100	600

Total costs Thousand Euro	Investment cost	Annual fixed maintenance costs	fuel cost
Technology C	500	300	750

Below is shown how discount rate can affect the analysis.



The main steps involved

A cost effectiveness analysis will normally involve four stages. Firstly, the programme objectives are determined. Then the total public sector resource costs of the programme are assessed. Generally, only direct monetary resources are included, although the programme costs may sometimes be measured in relation to the benefits that could have been obtained by allocating the monies to other projects (i.e. the opportunity cost). Thirdly, the impact is measured, with due assessment given to additionality (see below). Finally, the cost per unit output and outcome are assessed, through the simple division of costs by outputs/outcomes. Since it makes explicit the relationship between inputs and outputs, and thus the efficiency of the programme, it can provide useful insights into the programme.

Step 1.

First and foremost the expected result of the project must be identified and quantified in physical terms (e.g. number of road accidents avoided, number of new trained workers after a course). The following questions should be answered: What are the goals to be achieved? What are the programme outputs? What are the expected impacts? Which one of these may be considered predominant?

Step 2. Definition of total cost of the programme

The total cost of the intervention must be calculated. If possible, the basic rules of cost-benefit analysis can be applied to define costs. At this stage the cost of all the public resources of the programme are added up to obtain a total cost. Generally speaking, only direct resources that have a well-defined monetary value are included. The cost of a programme is sometimes measured in relation to the benefits that could have been obtained by allocating public monies to other ends (opportunity costs).

Step 3. Measuring the impact

This step is the trickiest one. Numerous studies use empirical methods based on the collection of primary data in order to gather information on the positive effects of a programme under evaluation. It is also possible, however, to estimate impacts on the basis of secondary data and/or the modelling of the implementation of the programme.

This should be the case for different alternatives with the same time horizons with different investment and recurrent costs and different level of the same benefit achieved during the entire life cycle of the project. How could this projects be compared? In this case an annual equivalent value of costs should be compared with the annual benefit level.

Whatever the method used, it is important to have an exact picture not only of the positive results , but also of deadweight losses, employment displacement effects, investment crowding out effects, etc. Certain evaluators have suggested that indications should be given about output multiplier and indirect or secondary results. In practice it is extremely difficult to accurately evaluate these secondary results. Consequently, the majority of evaluations do not take them into account. This solution is not commendable given the importance of secondary results in some circumstances. The lack of accuracy in the estimation of these impacts can be attenuated by the use of a sensitivity analysis.

Step 4. Calculation of the cost-effectiveness ratio

The last phase is that of the calculation which will give the final result. It consists of a simple ratio.

On the other hand the cost-effectiveness ratio should be used with caution. In the example on *How to calculate the cost-effectiveness ratio*, for instance, it would have been more correct to compare the ratios obtained in the case of different unit costs for the same expected benefit (e.g. unit costs of 300, 500, 200 and 100 for a result of 20), or different benefits for the same unit cost (e.g. results of 20, 12, 30 and 60 for the same cost of Euro 300). In other words, the cost-effectiveness ratio can be used as a single criterion of selection only in cases where the denominator or the numerator are respectively the same for each alternative.

How to calculate the cost-effectiveness ratio

Four project options to improve mathematical skills:

- *Small remedial groups with a special teacher*
- *A self-study programme supported by specially designed materials*
- *Computer-assisted learning*
- *A programme involving peer tutoring*

The expected output is measured by test scoring.

The table shows the calculation of the cost-effectiveness ratio for each option:

Intervention	*Size of effect on test scores*	*Cost per student (Euro)*	*Cost effect ratio*
Small groups with teacher	*20*	*300*	*15*
Self-study materials	*4*	*100*	*25*
Computer-assisted learning	*15*	*150*	*10*
Peer tutoring	*10*	*50*	*5*

The Peer Tutoring Project is the most cost-effective, Source: Belli, 2001.

Strengths and limitations of the technique

The analysis must be preceded by consideration of the programme objectives, of its main objective purpose and of the appropriate indicator applied to that objective. It provides an alternative to cost-benefit analysis, when outputs are not easy to monetise but can be quantified on a physical unit of account.

It is sometimes useful for evaluating the expected impacts in the ex-ante appraisal and for calculating the achieved impacts in the ex-post evaluation. By nature CEA tends to focus on direct results that occur over the short to medium term, but does not usually look at more long-term impacts. Yet it is on these that the overall effectiveness or the lack of effectiveness of programmes and policies crucially depend. For example, the effectiveness of Structural Fund programmes depends on achieving a combination of multiple objectives aimed at promoting growth and endogenous development within a region.

Comparisons (or benchmarking) of cost effectiveness ratios between projects are possible, but require considerable care. Often the approach used to assess costs or outputs may vary. Thus, one evaluation may express prices on a constant basis, while another uses nominal figures. There may also be qualitative differences in the outputs or in the results. For example, the remuneration or longevity of the net additional jobs created by two projects may be substantially different, although the same number of jobs are involved.

Cost effectiveness can be used in combination with other methods to data analysis. Stakeholder consultations, focus groups and expert panels could be a means of getting a wider understanding of the key issues in the overall social and economic context, which need to be taken into account in the value for money judgement.

Usually, the concept is dynamic in that it feeds into decisions about funding and allocations as the programme goes along. This is why assessing the cost effectiveness and the value for money of an intervention can be done ex ante, mid term and ex post. CEA can serve to compare programmes only when their implementation is straightforward and their impacts of a similar nature. Since CEA is based on an estimation of the programme impact in relation to its main objective, it presents the advantage of producing easily understood findings, which concentrate on the main concerns of key groups (including politicians and decision-makers). If the analysis is based on specific links between inputs and outputs that are relatively well established, the tool can facilitate the description of the actual functioning of programmes. This can be useful for refining existing policies or improving the effectiveness of future interventions.

CEA can only be used to compare programmes that are simple to implement and have the same type of impact. However, this situation is far more infrequent than one would expect. Thus, for example, even programmes that target an identical major objective, such as the creation of jobs, can create employment opportunities that are qualitatively different (e.g. in terms of longevity, security, remuneration, probability that they will be accessible for the inhabitants of targeted regions, etc.).

Establishing causal links of this type requires the ability to obtain detailed data from programme managers and addressees. When these data are not collected via the monitoring of programmes and projects, the evaluation will involve the painstaking collection of primary data. Ideally cost-effectiveness analysis should be used in combination with other techniques of economic analysis in order to analyse longer-term impacts on the regional GDP and on competitiveness.

Bibliography

Belli, P., Anderson, J. R., Barnum, H.N., Dixon, J. A., Tan, J-P, 2001, Economic Analysis, of Investment Operations. Analytical Tools and Practical Applications, WBI, World Bank, Washington D.C.

D. Potts, 2002, Project Planning and Analysis for Development, Lyann Rienner Publishers.

ODA, 1988, Appraisal of projects in developing countries, A Guide for Economists, London.

H.E. Freeman, P.H. Rossi, S.R. Wright, 1979, Evaluer des projets sociaux dans les pays en développement, Centre de développement de l'organisation de coopération et de développement économiques.

5. Delphi survey

Description of the technique

The Delphi Survey is based on a structured process for collecting and synthesising knowledge from a group of experts by means of a series of questionnaires accompanied by controlled opinion feedback (Adler and Ziglio, 1996). The questionnaires are presented in the form of an anonymous and iterative consultation procedure by means of surveys (postal and/or e-mail).

The Delphi Survey originated as part of a post-war movement towards forecasting the possible effects of technology development in relation to economic and social re-generation. The technology forecasting studies were initiated by the Douglas Aircraft Company, which established the RAND project in 1946 to study the "broad subject of inter-continental warfare" (Fowles, 1978). The theoretical and methodological basis for forecasting was elaborated in a subsequent series of papers produced by the project. These argued that, in the absence of an established evidence base, emergent fields of enquiry could begin to develop such an evidence base through capturing and synthesising the opinions of domain experts. The Delphi Survey was therefore an attempt to 'align' the sometimes conflicting positions of experts into a coherent and unified perspective.

The technique is relatively simple. It consists of a series of questionnaires sent to a pre-selected group of experts. These questionnaires are designed to elicit and develop individual responses to the task specified and to enable the experts to refine their views as the group's work progresses in accordance with the assigned task. The rationale behind the Delphi Survey is to address and overcome the disadvantages of traditional forms of 'consultation by committee', particularly those related to group dynamics.

The purpose of the technique

Delphi is primarily used to facilitate the formation of a group judgement (Helmer, 1977). It developed in response to problems associated with conventional group opinion assessment techniques, such as Focus Groups, which can create problems of response bias due to the dominance of powerful opinion-leaders (Wissema, 1982). It may be used in forward planning to establish hypotheses about how scenarios are likely to develop, and on their socio-economic implications. For example, it has been widely used to generate forecasts in technology, education, and other fields (Cornish, 1977). Fundamentally, the method serves to shed light on the evolution of a situation, to identify priorities or to draw up prospective scenarios.

Circumstances in which it is applied

Although the method was originally developed to capture expertise in uncertain and emergent domains, it tends to be used in evaluation when significant expertise exists on the subject, for example in the case of programmes that are not innovative. The method is recommended when the questions posed are simple (a programme with few objectives, of a technical nature) and for the purpose of establishing a quantitative estimation of the potential impacts of an isolated intervention (e.g. increase in taxes or in the price of energy). It is also recommended in an ex ante evaluation context if the evaluation concerns public intervention of a technical nature. Thus, it was very often used in the framework of energy policies, for example, for prospective studies on the impact of changes in taxation. In the case of the evaluation of Structural Funds, for example, the Delphi Survey has been recommended for obtaining macro-economic estimations when the phenomena involved are complex; for example, to quantify the impact of a major infrastructure project. It may also be used to specify relations of causes and potential effects in the case of innovative interventions. It is particularly useful when a very large territory is being dealt with since there are no experts' travel expenses, only communication costs.

It has found to be particularly useful in programmes related to public health issues (such as, policies for drug use reduction and prevention of AIDS/HIV) and education (Adler and Ziglio, 1996; Cornish, 1977). According to a number of commentators, context is everything in deciding whether and when to use the Delphi method. According to Adler and Ziglio (1996), the key questions that need to be asked are:

- What kind of group communication process is desirable in order to explore the issue?
- Who are the people with expertise on the issue and where are they located?
- What are the alternative techniques available and what results can reasonably be expected from their application?

The main steps involved

The method consists of questioning the experts by means of successive questionnaires, in order to reveal convergence and any consensus there may be. The main stages of this process are (Fowles, 1978):

Step 1. Determination and formulation of questions

Particular care must be given to the choice and formulation of questions, so as to obtain useful information.

Step 2. Selection of experts

They must have specific knowledge on the subject and be prepared to become involved in this type of procedure. The panel is generally composed of about fifty persons.

Step 3. Formulation of a first questionnaire that is sent to the experts

The first questionnaire must contain a reminder of the nature of the study and include two or three semi-open and open questions.

Step 4. Analysis of the answers to the first questionnaire

The answers are analysed in order to determine the general tendency and the most extreme answers.

Step 5. Formulation of a second questionnaire that is sent to experts

Each expert informed of the results of the first round is asked to provide a new answer and to justify it if it differs from the general tendency.

Step 6. Sending of a third questionnaire

This questionnaire is intended only for those experts whose answers were "extreme". They are asked to criticise the arguments of those who supported the opposite point of view. The comparison of opinions has a moderating influence and facilitates the appearance of convergence between the points of view.

Sufficient convergence of opinions generally appears with the fourth questionnaire. If that is not the case, the cycle continues.

Step 7: Summary of the process and drawing up of the final report.

It is important to note that the analysis of data elicited through Delphi surveys should be carried out using statistical analysis (for example cluster analysis of canonical correlation analysis) in order to identify convergences and divergences in responses.

Strengths and limitations of the technique

As has often been remarked, the results of a Delphi survey are only as valid as the opinions of the experts involved (Martino, 1978). Martino is only one of a number of critics of the Delphi method. The key problems reported include: poor internal consistency and reliability of judgements among experts, and therefore low reliability of forecasts based on the results elicited; sensitivity of results to ambiguity and respondent reactivity in the questionnaires used for data collection; difficulty in assessing the degree of expertise held by participating experts (Makridakis and Wheelright, 1978).

A major problem identified by research into the implementation and application of Delphi surveys has been the tendency for experts to over-simplify particular issues, and treat them as isolated events. This is particularly the case in forecasting, where experts tend to think in terms of linear, sequential events, rather than applying a holistic view that involves complex chains and associations. This has led to the development of techniques such as 'cross impact matrix forecasting', which are intended to compare a range of 'possible futures' against each other, and to consider the displacement, substitution and multiplier effects associated with the scenarios identified by the experts involved (Gordon and Hayward, 1968; Gatewood and Gatewood, 1983; Adler and Ziglio, 1996).

On the other hand, there have been several studies (Ament, 1970; Wissema, 1982; Helmer, 1983) supporting the Delphi method. These studies seem to suggest that in general, the Delphi method is useful to explore and unpack specific, single-dimension issues. There is less support for its use in complex, multi-dimensional modelling. In these cases, the evidence does suggest that data gathered by Delphi surveys is a useful input, when supported by data gathered from other sources, to complex scenario-building.

Bibliography

Nadeau M-A. (1988). L'évaluation de programme. Laval, Québec: Presses de l'université de Laval. pp 349-352.

A manual presenting the methods and instruments used in programme evaluation. The Delphi method is described briefly.

Godet M. (1985), Prospective et planification. Paris: Economica, pp. 119-125. The Delphi method is presented and illustrated in the first part of this book devoted to forecasting,

Witkin B.R et J.W Altschuld (1995), Planning Conducting Needs Assessments, Thousand Oaks: Sage. pp 193-203.

6. Expert panels

Description of the technique

An "expert panel" is a specially constituted work group that meets for evaluation. Expert panels are usually made up of independent specialists recognised in the fields covered by the evaluated programme in the evaluation process, usually as a mechanism for synthesising information from a range of sources, drawing on a range of viewpoints, in order to arrive at overall conclusions. To some extent, the expert panel draws largely upon legal practices in that results are usually based on reaching a consensus of opinion. Expert panels are a means of arriving at a value judgement on the intervention and its effects, which incorporates the main information available on the intervention, as well as previous and external experiences.

The panel may be considered as an evaluation tool in so far as there is a standard and reproducible procedure for forming it, bringing it together and leading it to produce its conclusions. Inspiration for the method was based on university juries - which explains why it appeared in the early 1970s - in the field of Research and Development evaluation. (The Delphi survey technique also relies on experts but differs in several other respects).

The experts are chosen to represent all points of view, in a balanced and impartial way. These experts are independent specialists, recognised in the domain of the evaluated intervention. They are asked to examine all the data and all the analyses made during the evaluation, and then to highlight consensus on the conclusions that the evaluation must draw, and particularly on the answers to give to evaluative questions. The panel does not fully explain its judgement references nor its trade-off between criteria, but the credibility of the evaluation is guaranteed by the fact that the conclusions result from consensus between people who are renowned specialists and represent the different "schools of expertise".

The purpose of the technique

The expert panel is used mainly to assess a programme or intervention, but it is a generic tool. The terms of reference given to the panel may include a wide range of questions, from the relevance of programme objectives to an estimation of real or probable effects.

Expert panels can be particularly helpful in arriving a judgements relating to quality and relevance. This tool is used for research programmes or innovation ones such as clusters.

Expert panels are also useful in the process of estimating impacts, especially to provide an interpretation and development of findings from evaluation work using other techniques

Circumstances in which it is applied

The tool is recommended when sufficient expertise exists in the field and when the evaluation is complex.

Expert panels are used to reach consensus on complex and ill-structured questions for which other tools do not provide univocal or credible answers. It is a particularly useful tool in relation to complex programmes, when it seems too difficult or complicated, in an evaluation, to embark on explanations or the grading of criteria in order to formulate conclusions.

It is also well suited to small, simple programmes, the evaluation of which does not warrant the mobilisation of many resources. The use of groups of experts makes it possible, within a few months, to gather the main points of view and knowledge relevant to the evaluation.

In relation to Structural Funds programmes, the expert panel technique could be used to carry out ex ante evaluation (although this is likely to take some time). For example, an expert panel could be asked to estimate the probable impact of a programme in terms of employment, or to assess the merits of the programme in terms of potential synergies. Expert panels are likely to be particularly useful in estimating probable impacts when used in conjunction with micro and macro economic modelling techniques, and indeed are a good way to judge whether the effects are sufficient or insufficient.

In the domain of European cohesion policy, expert panels can help to draw conclusions on the impacts if programmes which are not directly comparable, for example, synthesising qualitative conclusions. Sometimes formalised scoring systems may be used to bring together the views of the experts to arrive at a conclusion.

The expert panel may be used to formulate an independent, authoritative judgement, which is particularly useful in a partnership context, especially if there are differences in the partners' views.

As can be seen from the range of usages above, the technique is extremely versatile, and can be useful every time the structuring or judgement stage needs to be reinforced. For instance, the expert panel may intervene at the beginning and end of the evaluation, in combination with other tools used for the collection or analysis of the data.

Expert panels may fulfil numerous functions, it is preferable however to limit its work to only a part of the evaluation: the structuring of objectives and estimations of effects or judgements. The more clearly the panels' work is defined, the more its significance will be recognised. The reliability of the tool may be undermined if the questions put to the experts are too broad.

The main steps involved

Step 1. Identification of a list of potential experts

The members of the panel must be specialists recognised in at least one of the fields concerned by the programme. They must have extensive experience in the field and be independent vis-à-vis the commissioners of the evaluation. They must also have the required availability and wish to become involved in the evaluation.

The risk of bias from empathy is significant in so far as panels are too often limited to the specialists in the fields covered by the programmes (i.e. peers) who are hardly inclined to criticise the relevance of the objectives or to focus on any undesirable effects.

In the context of partnership programmes, it is possible to ask each partner to choose some of the experts, based on the similarity of their points of view. A list of experts exceeding the needs of the panel may also be proposed, and each partner be given the right to delete one or two names from the list.

The experts are nominated *intuitu personae* and do not represent their institution. Each expert signs a contract that, depending on the case, does or does not provide for remuneration.

Step 2. Selection and mandating of the experts

The panel is generally composed of between six and twelve members belonging to different "fields of expertise". The current tendency is to broaden the range of interests and to seek the greatest possible diversity of points of view in the panel.

The Chairperson for the panel is chosen by the commissioner or elected by her or his peers. It is essential that the secretariat of the panel be entrusted to a person with sufficient availability, which is generally not the case with the experts themselves.

Step 3. Investigations

The experts meet between three and six times, at one-month intervals. All the dates of their meetings must be planned at the outset. The panel's internal debates are under the seal of secrecy.

The members of the panel consult the programme or project documents (reports, preliminary studies, inquiries) and interview the programme leaders and several typical addressees. They may also make field visits, generally in groups of two to limit the risks of bias.

The tool is more likely to produce creative conclusions and rich recommendations if it is combined with suitable leadership methodologies such as METAPLAN or colour vote.

It is unlikely, apart from in the case of small and simple programmes, that the expert panel would be the only tool used to evaluate.

Indeed, expert panels tend to be used in conjunction with other information collection methods and analysis.

Step 4. Synthesis

The panel produces a report and formulates conclusions and recommendations that are collectively accepted. In case of disagreement, it may be useful to express the majority conclusion and to attach a comment by the minority expert.

Strengths and limitations of the technique

The expert panel is a very flexible tool that can be used to produce a synthetic judgement based on qualitative and quantitative data, even if these are incomplete. Its conclusions enjoy a high degree of credibility when recognised experts are used.

The expert panel is a relatively inexpensive and rapid tool.

The expert panel constructs a synthetic judgement of the programme being evaluated. This tool, when implemented with optimum efficiency, enhances the credibility and acceptability of the evaluation conclusions because differences between points of view are respected and consensus is reached. For partnership programmes, with this tool any differences between the points of view of the partners can be taken into account.

However, there are potential weaknesses. The experts must have extensive experience in the field, and therefore are at risk of bias and unwillingness to criticise the relevance of the objectives or to focus on any undesirable effects. Moreover, the comparison of opinions often leads to the under-evaluation of minority points of view. The consensual mode of functioning on which the dynamics of the panel is based, produces a convergence of opinions around majority values which are not necessarily the most relevant.

To some extent the potential weaknesses of expert panels can be avoided by taking precautions in the way they are assembled and organised. This could include:

- limiting its work to only a part of the evaluation: the structuring of objectives and estimations of effects or judgements, in order to ensure a clear focus and that its significance will be recognised;
- having a broad range of interests represented, including independent experts who are objective.

Bibliography

Nadeau M-A. (1988), L'évaluation de programme. Laval, Québec: Presses de l'université de Laval. pp. 349-352: This manual presents the methods and tools of programme evaluation; a few pages are devoted to expert panels.

Witkin B.R et J.W Altschuld (1995). Planning Conducting Needs Assessments, Thousand Oaks: Sage. pp. 193-203: This work presents a series of tools applicable to the evaluation of needs; a section is devoted to expert panels.

Callon M., Laredo P.et P. Mustar (1995), La gestion stratgique de la recherche et de la technologie. Paris: Economica. pp. 31-88: Presentation in the first two chapters of this book of the expert panel model used by the Commission of European Communities.

Cozzens S.E. (1987), 'Expert Review in Evaluating Programs', Science and Public Policy, 14(2), 71-81: Inventory and analysis of the American experience of the use of expert panels for evaluation. The interest of the conclusions transcends the context of science policy.

7. Focus groups

Description of the technique

The focus group is a well-established method of social inquiry, taking the form of structured discussion that involves the progressive sharing and refinement of participants' views and ideas. First used in market research, it is now applied widely in a variety of application and academic research settings to generate data and insights. The technique is particularly valuable for analysing themes or fields which give rise to divergent opinions or which involve complex issues that need to be explored in depth.

The focus group is one of a family of group based discussion methods. The typical format involves a relatively homogenous group of around six to eight people who meet once, for a period of around an hour and a half to two hours. The group interaction is facilitated by the evaluator or researcher who supplies the topics or questions for discussion. A variation is the workshop, implying a larger group, meeting in a larger session, with a more structured agenda. Other innovative methodologies involve the application of discussion group approach to decision-making. These include, for example, citizens' juries which bring together groups of between 12 and 30 people over the course of several days. They hear from 'witnesses', deliberate, and make recommendations about courses of action. Variations of this consultative approach include deliberative polls and consultative panels. The common features of these approaches are that they combine opportunities for accessing information with discussion and deliberation.

Although focus groups and other kinds of group-based discussions usually involve a physical coming together of participants, there is a growing interest in virtual groups that exploit advances in new information and communication technologies. The conduct of telephone groups using teleconferencing technology has in recent times been supplemented by online focus groups, involving web-mediated synchronous and asychronous discussion. The Delphi technique can also be readily adapted to electronic communication, although it does not feature truly interactive exchange. Here, views are gathered from group members individually and then summarised and circulated for further discussion until consensus is reached.

The purpose of the technique

The focus group makes it possible to bring together, simultaneously or sequentially, the different stakeholders in a programme (managers, operational staff, recipients or beneficiaries of services), and to collect a large amount of qualitative information in a relatively short space of time. In sharing and comparing their experiences and views, participants generate new insights and understandings. The method enables the evaluator to examine participants' different perspectives as these are constructed by their participation within a social network, and to explore how accounts are shaped through conversation with others in a naturalistic group context.

By playing on the interaction and confrontation of different points of view, the technique serves to reveal the participants' perceptions and views on topics and questions relevant to the evaluation. These may relate to its implementation, outputs or results. The facilitating role of the evaluator in the focus group discussion is aimed at opening out discussion and widening the range or response. Participants are encouraged to take the conversation into new and often unexpected directions, opening up different angles on evaluation topics and probing at deeper levels. Focus groups may also provide an effective means of evaluating sensitive topics.

Focus groups are a form of participatory evaluation. By involving the actors or beneficiaries of a programme as co-participants, the conclusions of the study will be more credible and more readily accepted. The focus group technique may also be used for the validation of data collection, or for complementing quantitative data.

Circumstances in which it is applied

Focus groups are a primary source of qualitative data, commonly combined with other qualitative methods and incorporated into a case study. Focus groups are well adapted to those cases where the evaluation topics and issues to be addressed provoke divergent opinions but where discussion may lead to a deeper and more considered viewpoint.

Focus groups are a valuable method in programmes where there is a power differential between participants and decision-makers. Current best practice is to work with homogeneous but contrasting groups, thereby producing information that can illuminate the distinctive perspectives, experiences and views of different stakeholders in the evaluation.

There is potential for application of focus groups to the different kinds of Structural Fund evaluations. The focus group technique may be used to test an innovative measure (ex ante evaluation), clarify the objectives of a project, establish a theory of change for the programme being evaluated, and identify the problems and needs of a region and the improvements required during the implementation of the programme. The technique is also very relevant at the end of the programme in the framework of ex post evaluation, to collect information for identifying and/or interpreting the results of the programme concerned, and to fix new priorities and orientations.

The main steps involved

Step 1. Selection of participants

The composition of the group, and the number of focus groups, depends on the particular requirements of the evaluation. It is preferable to select participants so as to ensure there is a degree of homogeneity in the group, and to form several groups of different composition. Limiting the work to a single group may undermine the validity of the study. It is unhelpful if there are significant imbalances in social power or status within the group. Diversity in other characteristics represented within each focus group is however desirable. The optimal number of participants is around 6 to 8 per group, so that each person has a turn to speak, and so that sub-groups are not formed. Usually the participants do not know each other as this facilitates both open questioning and disclosure. However, in some circumstances it can be beneficial to work with a naturally occurring group. For example, where the topic for evaluation concerns how an organisation understands a policy objective and how this translates into practice.

The participants are sometimes remunerated. This may be an incentive remuneration, or may be the offer to refund their travel expenses or to serve refreshments at the end of the session.

Step 2. Choice and training of facilitators

The facilitator's role is critical to the success of the group discussion. It requires good group facilitation skills and qualities to put people at their ease, to project oneself in positive ways to encourage a group, and to hold the interest of participants through to the end. Facilitators need good communication skills, a sensitivity to the issues under discussion, and a capacity to probe a topic to achieve greater depth as well as to challenge apparent consensus where this is led by conformity to social norms. It may be necessary to select facilitators who already possess these skills, or to institute some kind of training of evaluation staff to develop these group skills.

It is useful to make provision for a second person per group, particularly if the session is not tape recorded, so that one person can take notes while the other leads and facilitates the discussion. Co-facilitators may observe the discussion and then make recommendations to the facilitator on the way the meeting was conducted.

Step 3. Defining the interview topics

It is important to carefully define and limit the topics addressed as all participants must have an opportunity to participate in the discussion. A list of four or five open questions expressed in simple language is usually sufficient for a normal focus group session. The questions must be carefully defined, and arranged in a series with the most general ones first. The evaluator's aim is to use the opening question or topic to engage as many of the participants as possible and to promote discussion.

Step 4. Course of the discussion

The discussion may be launched fairly openly by introducing the subject of the session and asking a simple question of general interest. This will enable each participant to give an initial opinion or remark on the subject. As the discussion moves on, the aim is to clarify, delve deeper and to cover all angles. The facilitator's aim is to allow as much relevant discussion as possible to be generated from within the group, while at the same time ensuring that the topics and questions of interest to the evaluation are covered within the allotted time. This involves deciding when to move the discussion on to another topic, keeping the discussion relevant and focused, and choosing when to allow more free-ranging discussion with minimal intervention. In orchestrating the flow of contributions, a combination of assertiveness and tact may be required.

Step 5. Analysis and report on results

This final phase consists of interpreting and comparing the information given by the participants, and looking for shared and divergent opinions within each group. The information collected is codified so as to organise the results in relation to the objectives of the evaluation. The interpretation of data must take into account and distinguish two major aspects of the discussion: what the participants consider as interesting, and what they judge as being important. The analysis will depend on the number of focus groups questioned, and on the nature of the interviews (for example, did the focus group discussion take a structured approach, or not?). The results from the different groups are compared so as to identify any convergence there may be. The report may quote the most noteworthy statements made by the participants, together with a summary of the discussion.

Strengths and limitations of the technique

This type of discussion group method provides in-depth information on the values and opinions of selected participants. As the data emerges from discussion within the group, the perspective is less influenced by interaction with the researcher than it might be in a one-to-one interview.

The fact of bringing a number of people together provides a certain balance in the answers given and makes it easier for the evaluation team to define the general opinion on a particular programme. Owing to the participation of several persons, the focus group provides a level of 'quality control' over data collection by judging the pros and cons of each person's arguments and thus avoiding extreme opinions.

In a short space of time (from one and half to two hours), it is possible to collect a large amount of qualitative information.

Specific skills are required for managing the group dynamic and obtaining a balanced discussion while avoiding the dominant influence of opinion leaders in the group.

The discussion may sometimes be biased, due to the fact that the participants (beneficiaries) of public policies are subject to an effect of dependency and will produce a positive judgement a priori. An opposite dynamic sometimes observed in groups, especially in situations where there are few opportunities to voice opinions, is for programme participants to dump their frustrations about some new policy initiative.

It is possible that participation in a focus group changes peoples' perceptions - either because of the 'Hawthorne effect' (the fact that the behaviour of persons who know themselves to be under observation changes) or because their interaction with other participants gives them new insights and perspectives. Thus for example programme managers may actually improve their performance as a result of participating in a focus group. In this way the focus group methodology may therefore have an impact on the programme being evaluated or on successor programmes. The focus group thus becomes a form of action learning.

Bibliography

Anzieu, D. and Martin, J.Y. (1994) "La dynamique des groupes restreints", PUF Le psychologue. Presentation by psycho-sociologissts of the concept of a limited group, the main group phenomena and some fields of application.

Aubel, J. (1992) "Guide pour des etudes utilisant les discussions de group", BIT. Practical guide to conducting group interviews.

Finch, H. and Lewis, J. (2003) 'Focus groups'. eds: Ritchie, J. and Lewis, J. Qualitative Research Practice. London: Sage Publications.

Greenbaum, T.L. (1998) The Handbook of Focus Group Research. London: Sage Publications (second edition).

Kruger, R. (1995), Focus Groups: A Practical Guide for Applied Research, Thousand Oaks, CA; Sage Publications.

Morgan, D. L. (1997) 'Focus groups as qualitative research' in Qualitative Research Methods Series, 16. London: Sage Publications.

8. Impact evaluation[11]

An introduction illustrates the two conceptually distinct sets of questions behind impact evaluation. The first set of questions is devoted to establishing the theory behind an intervention and assessing whether it has been implemented according to that theory in order to judge the contribution of the intervention to observed effects, *theory based evaluation*. The second set of questions focuses on whether a given intervention produces the desired effects on some dimension of interest. The key question here: "does it make a difference?" This is answered by identifying and estimating causal effects through *counterfactual evaluation*.

The second chapter aims to explain approaches to theory based evaluation developed since the 1990s The third chapter aims to explain the counterfactual logic and its limitations: it introduces randomized controlled trials (so-called "experimental methods"), clearly stating their limited relevance within the evaluation of cohesion policy; finally, it explains the logic of non-experimental methods and how they deal with the problem of selection bias as well as natural dynamics bias. It then explains in accessible language the main counterfactual methods for estimating impacts, namely *statistical matching based on the propensity score*, *difference-in-differences, discontinuity designs and instrumental variables.*

Introduction to Impact evaluation

Quantifying and explaining the effects of interventions is at the heart of the evaluation of socio-economic development programmes. For policy makers to make informed decisions, it is important to understand what works or what does not, as well as why, for whom and in which contexts. This is a *formidable* list of questions, and the available analytical methods provide at best tentative and incomplete answers to most of them. Thus it is of fundamental importance to clarify which methods can answer which questions, under which circumstances.

Two distinct sets of questions (and methods)

Two conceptually distinct sets of questions tend to emerge when it comes to assessing the effects of public policies: one deals primarily with the quantification of effects, the other with their explanation.

- Methods *primarily* devoted to understanding why an intervention produces intended and unintended effects, for whom and in which context. The goal is to answer the "*why it works?"* question by identifying the theory of change behind the programme and assessing its success by comparing theory with actual implementation.

- Methods *primarily* devoted to establishing whether a given intervention produces the desired effects on some pre-established dimension of interest. The overarching goal is to answer a "*does it make a difference*?" question by identifying and estimating causal effects through counterfactual methods.

We want to stress the term "primarily". Identifying and estimating causal effects requires *some* theory, while comparing theory and implementation requires *some* quantification. However, these remain two distinct questions. It would be counterproductive, at this stage of the development and utilization of these methods, to force a synthesis between the two sets of questions and related methods.

[11] This section is based on guidance developed for DG REGIO on theory based impact evaluation by Professor Frans Leeuw of Maastricht University and on counterfactual impact evaluation by Professor Alberto Martini, Prova

Claims of cognitive superiority vs. intellectual honesty

Clear cut separation should help prevent antagonism, which is rife when proponents of alternative methods vie for the attention of the same policy makers and compete for the same resources. Claims of the alleged intellectual superiority of a set of methods over the other is the most deleterious manifestation of such rivalry and should be discouraged by openly rewarding the *opposite* attitude: the intellectually honest admission of the drawbacks, limitations and pitfalls of the analytical tools each side is able to deploy in answering questions about the what and why of the effectiveness of policy. Rhetorical claims of cognitive superiority should be left to the bygone era of the fruitless "paradigm wars". What the two camps mostly have in common is how little they truly understand about the effects of public policies.

While they should be kept separate methodologically, policymakers should use the results of both sets of methods as they see fit: "Even assuming that the counterfactual methods proved that a certain intervention worked and could even put a number on this, this is still a finding about one intervention under certain circumstances. We will need our more qualitative, "traditional" evaluation techniques to understand to which interventions these findings can be transferred and what determines the degree of transferability*" (*Stryczynski, 2009*).* Joint utilization is up to the user of the information, but it does not imply joint production.

Counterfactual impact evaluation (CIE) vs. Theory-based impact evaluation (TBIE)

The central question of CIE is rather narrow—how much difference does a treatment make—and produces answers that are typically *numbers*, or more often *differences*, to which it is plausible to give a *causal interpretation* based on empirical evidence and some assumptions. Is the difference *observed* in the outcome after the implementation of the intervention *caused* by the intervention itself, or by something else? Answering this question in a credible way is nevertheless a very challenging task.

The CIE approach to evaluation is useful for many policy decisions, because: (i) it gives easily interpretable information; (ii) it is an essential ingredient for cost-benefit and cost-effectiveness calculations; (iii) it can be broken down into separate numbers for subgroups, provided that the subgroups are defined in advance.

Howard White (2009), an advocate of TBIE, recognizes the importance of the following aspects of CIE: "*Criticisms of reporting an average treatment effect should not be overstated. Heterogeneity matters, as does understanding the context in which a particular impact has occurred. But it will rarely be the case that the average treatment effect (usually both the treatment of the treated and the intention to treat) is not of interest. Indeed it is very likely to be the main parameter of interest. It would be misleading to report significance, or not, a particular sub-group if the average treatment effect had the opposite sign. Moreover the average treatment effect is the basis for cost effectiveness calculations".*

To sum up, "how much difference does a treatment make" is an important, relevant, methodologically sound evaluation question. Yet it remains extremely challenging to answer, as the chapters on the various CIE approaches will openly document. But it is certainly *not* the only question.

The importance of TBIE stems from the fact that a great deal of other information, besides quantifiable causal effect, is useful to policy makers to make decisions and to be accountable to citizens. The question of why a set of interventions produces effects, intended as well as unintended, for whom and in which context, is as relevant, important, and equally challenging, if not more, than the "made a difference" question.

This approach does not produce a number, *it produces a narrative*. Thus it cannot be used for cost-benefit calculations, it is not communicated as quickly and schematically, and it is not backed by a comparable set of statistical tools. Thus it appears to some observers less scientific, less “objective”. But it can provide a precious and rare commodity, *insights* into why things work, or don’t. Above all, it is based on the very powerful idea that the essential ingredient is not a counterfactual (“how things would have been without”) rather a *theory of change* (“how things should logically work to produce the desired change”). The centrality of the theory of change justifies calling this approach *theory-based* impact evaluation.

Attribution vs. contribution

Causal questions are those that “strive to understand and assess relations of cause and effect (how and to what extent is what occurred attributable to the programme?)”. Thus, this notion of causality is centred on the idea of “attribution”. Causal questions related to the attribution of programme impacts appear frequently in the context of socio-economic development policy. For example, does aid to small and medium enterprises increase their survival or alter their hiring practices? Does investment in a new transport infrastructure eliminate bottlenecks and reduce travelling times? The ultimate objective in asking these questions is to learn whether the intervention works; which interventions produce the desired effect? Or, as seen from a different perspective, to what extent are the observed changes truly caused by the intervention?

In TBIE, causality is often declined as a problem of *contribution, not attribution*. Often quoted is *causal contribution analysis* (Mayne, 2001; Leeuw, 2003) which aims to demonstrate whether or not the evaluated intervention is one of the causes of observed change. Contribution analysis relies upon chains of logical arguments that are verified through a careful field work. Rigour in causal contribution analysis involves systematically identifying and investigating alternative explanations for observed impacts.[1] This includes being able to rule out implementation failure as an explanation of lack of results, and developing testable hypotheses and predictions to identify the conditions under which interventions contribute to specific impacts.

It is not “complexity” driving the difference...

A common perception is that the “*counterfactual impact evaluation*” is suited for “simple” intervention, while “*theory-based impact evaluation*” is necessary for complex intervention. This “division of labour” is by and large a misconception. First simple projects are components of complex intervention and the "Does it work" can be one element or a broader evaluation taking into account this complexity. Second, the “why it works” question is relevant also for relatively simple projects characterized by single ‘strand’ initiatives with explicit objectives. Actually, the “why it works” question might stand a better chance of finding an answer in these situations than in comprehensive programmes with an extensive range and scope, with a variety of activities that cut across sectors, themes and geographic areas. The very idea that complex situations are *easily* understood by complex methods is simply wrong: complexity is a problem for all.

...it is rather the disciplines lurking behind

The CIE approach to evaluation is backed by a formidable stock of methodological tools. The statistical/econometric/epidemiological community has produced in the last three decades a rather sophisticated conceptual apparatus to deal with causal inference: the potential outcome or *counterfactual approach*.

Quantifying effects requires establishing a counterfactual. That is to say, to reconstruct *what would have happened in the absence of the intervention.* This apparently simple idea turns out to be very powerful.

Section 8.4 is devoted to methods needed to answer this type of questions. The logic of causal explanation adopted by these methods is referred to as “counterfactual logic”. Its centerpiece is the notion of *causal effect* of an intervention, defined as the *difference* between the outcome observed *after* an intervention has taken place, and the outcome that would have occurred in *the absence* of the intervention: the latter is not observed and must be recovered from other data.

On the other hand, the field of theory-based impact evaluation is not lacking in the number of proposed methods. The literature on TBIE methodology is riddled with labels representing different (and sometimes not so different) methodological approaches. TBIE is backed by a vast array of qualitative, naturalistic, participatory, hermeneutic methods. However, these have not developed into a powerful and validated set of tools the CIE can draw upon.

Perhaps the most visible approach is *Realist evaluation* (Pawson and Tilley 1997; Pawson, 2002) that has spent a considerable amount of energy stressing the epistemological differences from CIE, proposing a different understanding of causality, based on a “*generative”* notion centred on the identification of causal mechanisms, rather than a mere “*successionist*” view, typical of the counterfactual approach. The basic idea of Realist evaluation is that different contexts may yield different reactions to the same intervention, and putting in place alternative mechanisms may produce different results.

Another example is the GTZ Impact Model, developed by the International Fund for Agricultural Development (IFAD), a specialized agency of the United Nations, which shows an ‘attribution gap’ between the direct benefits (which can be demonstrated through project level monitoring and evaluation) and the indirect, longer term development results (observed changes) of the intervention. Impact pathway evaluation represents a set of hypotheses about what needs to happen for the intervention outputs to be transformed, over time, into impact on highly aggregated development indicators.

Finally, *participatory evaluation* approaches are built on the principle that stakeholders should be involved in some or all stages of the evaluation. In the case of impact evaluation this includes aspects such as the determination of objectives, indicators to be taken into account, as well as stakeholder participation in data collection and analysis.

Selected references

Leeuw F. [2003], Reconstructing Program Theories: Methods Available and Problems to be Solved, in «American Journal of Evaluation», n. 24(1), pp. 5-20.

Leeuw F., Vaessen J. [2009], Impact Evaluations and Development: NONIE guidance on impact evaluation, Network of Networks on Impact Evaluation (NONIE).

Mayne J. [2001], Addressing Attribution through Contribution Analysis: Using Performance Measures Sensibly, in «Canadian Journal of Program Evaluation», n. 16(1), pp. 1-24.

Pawson R. [2002], Evidence-based Policy: The Promise of ‘Realist Synthesis’, in «Evaluation», n. 8(3), pp. 340-358.

Pawson R. and Tilley N. [1997], Realistic Evaluation, Sage Publications, Thousand Oaks, CA.

Riché M. [2013],Theory-based evaluation: a wealth of approaches and an untapped potential, in "Capturing effects of projects and programmes", pp.64-79, Studentlitteratur AB, Lund

Stryczynski K. [2009], Rigorous impact evaluation: the magic bullet for evaluation of Cohesion Policy?, European Commission, Bruxelles.

White H. [2009], Theory-Based Impact Evaluation: Principles And Practice, Working Paper n. 3, International Initiative for Impact Evaluation, New Delhi.

[1] A radical point of departure between the two approaches is the very concept of “observed impacts”. CIE contends that impacts are not observable, being the difference between something observable and a hypothetical state. TBIE, somehow less clearly, contends that impacts can be observed.

Theory-based Impact evaluation

Introduction to TBIE

Over the last forty years a number of evaluation experts (Suchman, 1967; Chen and Rossi, 1980; Weiss, 1995; Pawson and Tilley, 1997; Rogers et al., 2008; Donaldson, 2007) have contributed to - the development of what can be called a theory-oriented evaluation approach - called theory-based evaluation, theory-driven evaluation or programme theory evaluation. For the purposes of EVALSED, the term theory-based evaluation (TBE) is used to reflect these approaches.

The objective of this guidance is to provide users of EVALSED with some general ideas of what TBE is, what questions it can answer under which circumstances and how the approach can be applied, using various evaluations methods.

Several approaches have been developed within TBE over the years. However, these approaches have not been applied often within the socio-economic development programmes financed under EU Cohesion policy. Therefore, the present guidance provides examples of how TBE has been used in other intervention fields. When good practice examples are available in the field of the EU Cohesion policy, the guidance material will be updated.

Some of the data collection techniques relevant for TBE, e.g., focus groups, workshops, case studies, expert judgements, are explained elsewhere in this sourcebook.

Users of the guidance are encouraged to refer to the examples and references given before applying the approaches for the first time. The approaches build upon a wealth of experience and literature that is not fully reviewed here. Therefore, it would be advisable, especially for those who wish to explore TBE in more detail, to refer to additional expertise and other specialised literature.

Definition of Theory Based Evaluation

Theory-based evaluation is an approach in which attention is paid to theories of policy makers, programme managers or other stakeholders, i.e., collections of assumptions, and hypotheses - empirically testable - that are logically linked together.

These theories can express an intervention logic of a policy: policy actions, by allocating (spending) certain financial resources (the inputs) aim to produce planned outputs through which intended results in terms of people’s well-being and progress are expected to be achieved. The actual results will depend both on policy effectiveness and on other factors affecting results, including the context. An essential element of policy effectiveness is the mechanisms that make the intervention work. Mechanisms are not the input-output-result chain, the logic model or statistical equations. They concern amongst others beliefs, desires, cognitions and other decision-making processes that influence behavioural choices and actions. Theory based evaluation explores the mechanisms which policy makers believe make the policy effective and compares these with research based evidence.

Theory-based evaluation focuses on this intervention theory; it aims to find and articulate this theory, to test it and to improve it, if necessary.

Theory-based evaluation has at its core two vital components. The first is conceptual, the second empirical. Conceptually, theory-based evaluations articulate a policy or programme theory. Empirically, theory-based evaluations seek to test this theory, to investigate whether, why or how policies or programmes cause intended or observed results.

Testing the theories can be done on the basis of existing or new data, both quantitative (experimental and non-experimental) and qualitative. TBE does not apply a hierarchy of research designs and methods; it does not favour any over any others, as long as they are rigorously applied. Their choice depends on the evaluation design and they should be selected if they are appropriate to answer the evaluation questions.

Theories underlying a policy or programme are often not directly visible or knowable to evaluators. They are often not explicitly expressed in official documents. Evaluators have to search for these theories – if they have not been concisely articulated - and explain them in a testable way. Then they have to test them. Below are briefly presented some approaches to how to do this. The list is not exhaustive.

Realist Evaluation

The term realist evaluation was coined by Ray Pawson and Nick Tilley in their book with the same title (1997). This methodological approach stresses the importance of CMO (Context, Mechanism, Outcomes) configurations basic to policies and programmes.

Let us take an example of a socio-economic development programme. Such a programme attempts to solve a problem – to create some kind of socio-economic change. The programme works by enabling stakeholders to make choices. But choice-making is always constrained by stakeholders' previous experiences, beliefs and attitudes, opportunities and access to resources.

Making and sustaining different choices requires a change in stakeholders' reasoning (for example, values, beliefs, attitudes or the logic they apply to a particular situation) and the resources (e.g. information, skills, material resources, financial support) available to them. This combination of reasoning and resources is what enables the programme to work and is known as a programme 'mechanism'. The programme works in different ways and sometimes for different people (that is, the programme can trigger different change mechanisms for different stakeholders).

The contexts in which the programme operates make a difference to the results it achieves. Programme contexts include features such as social, economic and political structures, organisational context, programme stakeholders, programme staffing, geographical and historical context and so on. Some factors in the context may enable particular mechanisms to be triggered. Other aspects of the context may prevent particular mechanisms from being triggered. There is always an interaction between context and mechanism and that interaction is what creates the programme's results: Context + Mechanism = Result.

Because programmes work differently in different contexts and through different change mechanisms, they cannot simply be replicated from one context to another and automatically achieve the same results. Knowledge about ‘what works for whom, in what contexts, and how’ are, however, portable.

Therefore, one of the tasks of evaluation is to learn more about ‘what works for whom’, ‘in which contexts particular programmes do and don’t work’, and ‘what mechanisms are triggered by what programmes in what contexts’.

A realist approach assumes that programmes are theories incarnate. That is, whenever a programme is implemented, it is testing a theory about what might cause change, even though that theory may not be explicit. One of the tasks of a realist evaluation is, therefore, to make the theories within a programme explicit, by developing clear hypotheses about how, and for whom, programmes might ‘work’. The implementation of the programme, and the evaluation of it, then tests those hypotheses. This means collecting data, quantitative and qualitative, not just about programme results, or the processes of programme implementation, but about the specific aspects of programme context that might impact on programme results, and about the specific mechanisms that might create change.

Pawson and Tilley also argue that a realist approach has particular implications for the design of an evaluation and the roles of those benefiting from a programme. For example, rather than comparing changes for participants who have benefitted from a programme with a group of people who have not (as is done in random control or non-experimental designs), a realist evaluation compares mechanisms and results within and between programmes. It may ask, for example, whether a programme works differently in different localities (and if so, how and why); or for different population groups (for example, men and women, or groups with differing socio-economic status). They argue that different stakeholders will have different information and understandings about how programmes are supposed to work and whether they in fact do. Data collection processes (interviews, focus groups, collection of administrative data, questionnaires and so on) should be constructed to collect the particular information that those stakeholder groups will have, and thereby, to refute or refine theories about how and for whom the programme works.

Pawson and Sridharan (2010) present the following methodological steps for a realistic evaluation of a programme:

- *Eliciting and surfacing the underlying programme theories*

The first point is that although programme theories 'are easily spotted', these theories are best elicited from their procreators and this may involve either:

Reading and closely analysing programme documentation, guidance, regulations, etc., on how the programme will achieve its ends, or

Interviews with programme architects, managers or practitioners on how their intervention will generate the desired change.

Programme theories normally flow quite readily from these interviews, though some related difficulties should be noted. The first, located at the political level, is a tendency to ambiguity in policy discourse. The second problem, located nearer the programme practice, occurs when the core theory is either seemingly so obvious or buried tacitly in the minds of the programme makers that it can fail to surface in the interview. In these situations, persuasion is sometimes needed to encourage practitioners to spell out how their actions worm their way into participants’ choices (Pawson and Tilley 1997).

- *Mapping and selecting the theories to put to research*

Having found a means to elicit the programme theories at work, the next stage is to begin to codify or map them.

An array of techniques is available for this task, known variously as concept mapping, logic modelling, system mapping, problem and solution trees, scenario building, configuration mapping, and so on. All try to render the process through which the programme achieves its ends, usually in diagrammatic form. These maps may identify the various causes of the problem, the administrative steps to take, the unfolding sequence of programme activities and inputs, the successive shifts in dispositions of participants, the progressive targeting of programme recipients, the sequence of intervention outputs and results.

- *Formalising the theories to put to test*

After eliciting, mapping and selecting programme theories, the time comes to formalise them. They need to be transformed into a propositional form, as hypotheses suitable for empirical research. Programme theories come to life as insights, brain waves, bright ideas, and informed guesses. Sometimes, they turn out to be wishful thinking and pipe dreams. What evaluation research requires, by contrast, are testable propositions.

- *Data collection and analysis*

Data collection and analysis follows an empirical research (qualitative and quantitative, including experimental and non-experimental techniques) in order to understand, test and refine policy or programme theories (regarding CMO).

Realist evaluators refer to ´interrogating the policy or programme theories'; they also stress that whilst it was appropriate to follow custom and refer to this stage as 'theory testing', these approaches prefer the term 'theory refinement'. The objective is not to accept or reject policy/programme theories. The mission is to improve them.

Theory of Change

Carol Weiss (1995) popularized the term 'theory of change'. She hypothesized that a key reason complex policies or programmes are so difficult to evaluate is that the assumptions that inspire them are poorly articulated. She argued that stakeholders of complex initiatives are typically unclear about how the change process will unfold and therefore pay little attention to the early and mid-term changes that need to happen in order for a longer term goal to be reached. The lack of clarity about the 'mini-steps' that must be taken to reach a long term result not only makes the task of evaluating a complex initiative challenging, but reduces the likelihood that all important factors related to the long term goal will be examined. Weiss defined theory of change as a way to describe the set of assumptions that explain both the mini-steps that lead to the long term goal and the connections between policy or programme activities and results that occur at each step of the way. She challenged designers of complex initiatives, such as EU programmes, to be specific about the theories of change guiding their work and suggested that doing so would improve their policies and would strengthen their ability to claim credit for outcomes that were predicted in their theory.

The following steps elicit the theory of change underlying a planned programme. A pre-condition is that the evaluator works collaboratively with a wide range of stakeholders.

Step 1: The focus is on the long-term vision of a programme and is likely to relate to a timescale that lies beyond its timeframe. Its aim should be closely linked to the existence of a local, regional or national problem. For example, a smoking cessation programme might have a long-term vision of eradicating inequalities in smoking prevalence by 2020.

Step 2: Having agreed the ultimate aim of the programme, stakeholders are encouraged to consider the necessary results that will be required by the end of the programme if such an aim is to be met in the longer term. Within a programme they might, for instance, anticipate a decrease in gap between the most and least deprived areas.

Steps 3 and 4: Stakeholders are asked to articulate the types of outputs and short-term results that will help them achieve the specified targets. These might include reductions in differential access to acceptable smoking cessation programmes. At this stage those involved with the programme consider the most appropriate activities or interventions required to bring about the required change. Different strategies of engagement might be used to target pregnant women, middle-aged men and young adolescents, for example.

Step 5: Finally, stakeholders consider the resources that can realistically be brought to bear on the planned interventions. These will include staff and organisational capacity, the existence of supportive networks and facilities as well as financial capability.

Following a collective and iterative process the resulting programme theory must fulfill certain criteria: it must be plausible, doable and testable. It needs to be articulated in such a way that it can be open to evaluation; this is only possible where there is a high degree of specificity concerning the desired outcomes. Only then, the theory of change that is elicited should be interrogated to ensure that the underlying logic is one that is acceptable to stakeholders either because of the existing evidence base or because it seems likely to be true in a normative sense. The evaluator then takes the programme map generated through this process and, using various data (qualitative and quantitative) collection techniques as relevant, monitors and analyses the unfolding of the programme in practice and integrates the findings.

Contribution Analysis

'Contribution analysis' is a performance measurement approach developed within the Office of the Canadian Auditor General in the 1990s and it aims to establish the contribution a programme makes to desired outcomes.

In practice, many evaluations identify whether or not an result has been achieved and, if it was, assume the programme can take credit for this. However, reporting on results and proving attribution are two different things. Attribution involves drawing causal links and explanatory conclusions between observed changes and specific interventions. Determining whether an result was caused by a programme or other external factors is difficult and expensive. However, demonstrating the contribution of a programme to result is crucial if the value of the programme is to be demonstrated and to enable decisions to be made about its future direction (Mayne 2001).

Rather than attempt to definitively causally link a programme to desired outcomes, contribution analysis seeks to provide plausible evidence that can reduce uncertainty regarding the difference a programme is making to observed outcomes (Mayne 2001).

The following description is based on Mayne (2010) where seven methodological steps are described that form a contribution analysis.

Step 1: Set out the cause-effect issue to be addressed

First, it is necessary to articulate cause and effect. The following questions should be posed:

- Would the expected intervention make a difference to the problem?
- What aspects of the intervention or the context would lead to a contribution being made?
- What would provide evidence that the intervention made a noticeable contribution?
- Is the expected contribution plausible given the nature of the intervention and the problem being addressed? If not, the value of further analysis needs to be reassessed.

Step 2: Develop the theory of change

Developing a prior or initial theory of change for the intervention is the second key step. Contribution analysis needs reasonably straightforward, not overly detailed results chains, especially at the outset. Refinements may be needed to further explore some aspects of the theory of change, but can be added later.

Step 3: Assess the resulting contribution story

At this point, it is useful to critically review the contribution story resulting from the developed theory of change, i.e.:

- To assess the logic of the links and test the plausibility of the assumptions in the theory of change: Are there any significant gaps in the theory? Can they be filled by refining the theory of change? If not, is it worth continuing?
- To identify where evidence is needed to strengthen the contribution story: Which links have little evidence? Which external factors are not well understood?
- To determine how much the theory of change is contested: Is it widely agreed? Are specific aspects contested? Are there several theories of change at play?

Step 4: Gather existing evidence on the theory of change

Before gathering new data and information, it is useful and cost-effective to look at the relevant existing data and information there is about the theory of change. The purpose is to provide empirical evidence for the contribution story: evidence on activities implemented, on observed results, on assumptions being realised and on relevant external factors.

At this point in the analysis, a theory of change for the intervention has been developed and the available evidence supporting the theory of change collected. The theory of change has to some extent been tested. The significant external factors have also been identified and available evidence gathered for them.

Step 5: Re-assess the contribution story and challenges to it

The theory of change can be critically assessed in light of the existing evidence:

- Which links in the theory of change are strong (strong logic, good evidence available supporting the assumptions, low risk and wide acceptance) and which are weak?
- How credible is the story overall? Does the pattern of outcomes and links between them validate the contribution chain?
- Do stakeholders agree with the contribution story developed?

- Is it likely that any of the external significant factors have had a noteworthy influence on the results observed?
- What are the main weaknesses in the story? Where would additional data or information be useful?

Step 6: Seek out additional empirical evidence

This is the step where the primary data gathering for the evaluation begins, informed by the previous steps e.g., step 5 has identified where additional evidence is needed.

- Evidence is gathered to strengthen the contribution story, using appropriate data gathering techniques, such as surveys and reviewing and analysing administrative data. There may be evidence on results occurring, on the validity of the assumptions and risks in the theory of change and on significant external factors.
- There may possibilities to use quantitative techniques (experimental and non-experimental designs) involving comparison groups that could be used to explore elements of the theory of change.
- From a theory-based perspective, several frequently used data gathering techniques can be strengthened:
    - *Key informant interviews* can both test the theory of change developed and elicit alternative theories of change the key informants might have, as well as discuss other influencing factors. Interviewees should be asked what on evidence they base their views.
    - *Focus groups* and *workshops* can explore a theory of change since there will be discussion about how different people see the intervention working. Alternative theories of change may emerge and other influencing factors can be identified. They can be used to develop a theory of change and as a way to identify evidence on the extent to which the theory of change has been realised in practice.
    - *Case studies* can be used in the same way. Case studies are powerful as a data gathering tool to help confirm or refute a theory of change, or the micro steps in a theory of change, showing that the theory is indeed plausible and not just based on unsupported beliefs.

Step 7: Revise and strengthen the contribution story

Now, the newly collected empirical evidence should be used to build a more credible contribution story with strengthened conclusions on the causal links in the theory of change. Contribution analysis works best as an iterative process. At this point, the analysis may return to Step 5 and reassess the strengths and weaknesses of the contribution story and decide if further analysis is useful or possible.

Policy Scientific Approach

The 'policy scientific approach' covers the following six steps (Leeuw, 2003):

Step 1: Identify behavioral mechanisms expected to solve the problem

Searching in formal and informal documents and in interview transcripts can elicit statements that indicate why it is believed necessary to solve the policy problem and what the goals are of the policy or programme under review. These statements point to mechanisms; these can be considered the ‘engines’ that drive the policies or programmes and are believed to make them effective.

Step 2: Statements that have the following form are especially relevant for detecting these mechanisms:

- ‘It is evident that *x . . .* will work’
- ‘In our opinion, the best way to go about this problem is to *. . .*’
- ‘The only way to solve this problem is to *. . .*’
- ‘Our institution’s *x* years of experience tells us that *. . .*’;

Step 3: Compile a survey of these statements and link the mechanisms with the goals of the programme or policy under review

Step 4: Reformulate these statements in conditional ‘if–then’ propositions or propositions of a similar structure (‘the more x, the less y’).

Step 5: Search for 'warrants' to identify missing links in or between different propositions through argumentation analysis.

Argumentation analysis is a standard tool in logic and philosophy. It describes a model for analysing chains of arguments and it helps to reconstruct and fill in argumentations. A central concept is the 'warrant', the 'because' part of an argument: it says that B follows from A because of a (generally) accepted principle. For example, ‘the organisation’s performance will not improve next year’ follows from ‘the performance of this organisation has not improved over the last 5 years,’ because of the principle, ‘past performance is the best predictor of future performance.’ The ‘because’ part of such an argument often is not made explicit. Consequently, these warrants must be inferred by the person performing the analysis

Step 6: Reformulate these 'warrants' in terms of conditional 'if–then' (or similar) propositions and draw a chart of the (mostly causal) links.

Step 7: Evaluate the validity of the propositions by looking into:

- the logical consistency of the set of propositions;
- their empirical content, that is, the extent to which the theory and, in particular, the assumed impact of the behavioral mechanisms correspond with the state of the art within the social/behavioral/economic sciences on these mechanisms.

Evaluating the reconstructed programme theory can be done in different ways. One is to confront (or juxtapose) different theories (like Carvalho & White, 2004, with regard to social funds). Another is to empirically test the programme theory by making use of primary or secondary data (triangulation), both qualitative and quantitative. A third possibility is to organise an iterative process of continuous refinement using stakeholder feedback and multiple data collection techniques and sources (in the realist tradition), while a fourth approach is to make use of already published reviews and synthesis studies. These can play a pivotal role in marshalling existing evidence to deepen the power and validity of a TBE, to contribute to future knowledge building and to meet the information needs of stakeholders. Visualisation or mapping software can help in this task.

There are several techniques for data collection and analysis, for example:

- Systematic reviews are syntheses of primary studies that, from an initial explicit statement of objectives, follow a transparent, systematic and replicable methodology of literature search, inclusion and exclusion of studies according to clear criteria, and extracting and synthesizing of information from the resulting body of knowledge.
- Meta-analyses quantitatively synthesize 'scores' for the impact of a similar set of interventions from a range of studies across different environments.
- Realist syntheses collect earlier research findings by placing the policy instrument or intervention that is evaluated in the context of other similar instruments and describe the intervention in terms of its context, social and behavioral mechanisms (what makes the intervention work) and outcomes.

Strategic Assessment Approach

Central in the Strategic Assessment Approach are four major stages: (1) group formation; (2) assumption surfacing; (3) dialectical debate; and (4) synthesis (Leeuw, 2003; Mason and Mitrof, 1980).

Stage 1 - Group Formation: The aim is to structure groups so that the productive operation of the later stages of the methodology is facilitated. A wide cross-section of individuals with an interest in the relevant policy question should be involved. They are divided into groups, care being taken to maximise convergence of viewpoints within groups and to maximise divergence of perspectives between groups.

Stage 2 – Assumption Surfacing: The different groups separately unearth the most significant assumptions that underpin their preferred policies or programmes. Two techniques assume importance in assisting this process.

The first, stakeholder analysis, asks each group to identify the key individuals or groups upon whom the success or failure of their preferred strategy would depend. This involves asking questions such as: Who is affected by the strategy? Who has an interest in it? Who can affect its adoption, execution, or implementation? And who cares about it? For the stakeholders identified, each group then lists what assumptions it is making about each of them in believing that its preferred strategy will succeed.

The second technique is assumption rating. Initially one should find and list the assumptions. This involves searching for statements about symptoms of the problem (that have to be solved through a policy or programme, distinguishing them from statements about causes of the problem). For each of the listed assumptions, each group asks itself two questions: (1) How important is this assumption in terms of its influence on the success or failure of the strategy? And (2) how certain are we that the assumption is justified? Here, in fact, the evaluation of the listed assumptions takes place, usually by using research reviews and similar documents. The results are recorded on a chart. Each group then is able to identify a number of key assumptions upon which the success of its strategy rests.

Stage 3 - Dialectical debate: The groups are brought back together and each group makes the best possible case to the others for its preferred strategy, while identifying its key assumptions. Only points of information are allowed from other groups at this time. There is then an open debate focusing on which assumptions are different between groups, which are rated differently, and which of the other groups' assumptions each group finds most troubling. Each group should develop a full understanding of the preferred strategies of the others and their key assumptions.

Stage 4 – Synthesis: An attempt to synthesise is then made. Assumptions are negotiated and modifications to key assumptions are made. Agreed assumptions are noted; they can form the basis for consensus around a new strategy that bridges the gap between the old strategies and goes beyond them. If no synthesis can be achieved, points of disagreement are noted and the question of what research might be done to resolve these differences is discussed.

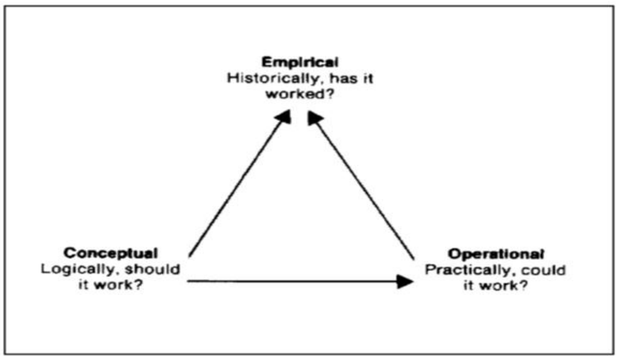
Prospective Evaluation Synthesis (PES) (GAO, 1995)

In essence, a prospective evaluation synthesis is a combination of: (1) a careful, skilled textual analysis of a proposed programme, designed to clarify the implied goals of the programme and what is assumed to obtain outcomes, (2) a review and synthesis of evaluations from similar programmes, and (3) summary judgments of likely success, given a future context that is not too different from the past. In this respect, the PES resembles the evaluation synthesis approach, except that the focus of the PES is on how evaluation studies cast light on the potential for success of the proposed programmes in the future, as opposed to reaching conclusions about the actual performance of existing programmes.

Conceptually, PES provides a way to use the logic of evaluation methodology and its procedures to assess the potential consequences either of one proposal or of alternative and competing policy proposals. It combines (1) the construction of underlying models of proposed programmes or actions as developed by Wholey for evaluability assessment with (2) the systematic application of existing knowledge as developed in the evaluation synthesis methodology (Wholey, 1977). PES is a prospective analysis anchored in evaluation concepts. It involves operational, conceptual, and empirical analyses, taken in the context of the future (see a figure below).

As the following figure illustrates, the conceptual analyses results help focus the operational analyses and answer the question: 'Logically, should the proposal work?' The operational analyses further scope the search for empirical findings and answer the question: 'Practically, could the proposal work?' The empirical analyses can open both new conceptual and operational possibilities and answer the question: 'Historically, have activities conceptually and operationally similar to the proposal worked in the past?' Finally, the PES takes into account ways in which the past is and is not likely to be similar to plausible future conditions.

Figure 3.1: the Triad of Analysis


Empirical
Historically, has it worked?
Conceptual
Logically, should it work?
Operational
Practically, could it work?

Elicitation Method

As policies and programmes are developed and implemented by organisations, the 'mental models' or 'cognitive maps' of people in these organisations, i.e., their theories, are important for understanding the anticipated impact of their policies or programmes. The emphasis should therefore be placed on organisational cognitions. One of the central questions is the relationships between these cognitions and the results of organisations. All stakeholders should have 'cognitions' (theories) about the organisation and its environment. These maps of what is going on in their organisation partly determine their behaviour. Their content concerns the organisational strategies, their chances of success, the role power plays, their own roles and the relationships with the outside world. Parts of these maps or theories are implicit and are tacit knowledge, both on an individual and on a collective level. By articulating these mental models, it is possible to compare them with evidence from scientific organisation studies. The articulation is also important for organisations to become 'learners'.

Examples of techniques for reconstructing, eliciting and assessing these mental or cognitive maps are the following:

- Look at the concrete record of strategic intentions, through, for example, a study of the documentation which is designed to direct behaviour;
- Look at decision-making in action; get involved in the organisation (an anthropological observer approach). Watch decision-makers, listen to stories;
- Work with managers on strategic breakdown situations. Become immersed in the thinking and the social process of 'strategic fire fighting';
- Use well-designed trigger questions in interview situations so that 'theories in use' can be detected. Follow interviews with feedback to individuals and to the team. The 'elicitation cycle' is built on responses to designed trigger questions. The process uses six techniques:
    - Create an open-ended atmosphere in the interview;
    - Do away with formal language and create a 'playful' atmosphere in which it is easier to deviate from the formal phraseology and the official script;
    - Do 'set the interviewees up against themselves';
    - Create dialectical tension by asking the interviewees to adopt unusual roles;
    - Listen very carefully for internal inconsistencies in what is being said;
- Apply data/content-analysis programmes or other text analysis programmes to the interview reports and documents; and
- Confront the results of these content-analysis activities with relevant (social) scientific research.

General Elimination Methodology, also known as Modus Operandi Approach

The core elements of the 'general elimination methodology' are the following (Scriven, 2008):

- The general premise is the deterministic principle: all macro events (or conditions, etc.) have a cause.
- The first 'premise from practice' is the List Of Possible Causes (LOPCs) of events of the type in which we are interested, e.g., learning gains, reduction of poverty, extension of life for AIDS patients. People have used LOPCs for more than a million years, in tracking and cooking and healing and repairing, and today every detective knows the list for murder, just as every competent mechanic knows the list for a brake failure, though the knowledge is as often tacit as explicit, outside the classroom and the maintenance videos. An LOPC usually refers to causes at a certain temporal or spatial remove from the effect, and at a certain level of conceptualisation, and will vary depending on these parameters; of course, the context of the investigation determines the appropriate distance parameters. The distant LOPC for murder is the list of possible motives; a more proximate one, developed in a particular case by applying the general one, is the list of suspects. When dealing with new effects, we may not be certain the list is complete, but we work with the list we have and extend it when necessary.

- The second practical premise is the list of the modus operandi for each of the possible causes (the MOL). Each cause has a set of footprints, a short one if it is a proximate cause, a long one if it is a remote one, but in general the modus operandi is a sequence of intermediate or concurrent events or a set of conditions, or a chain of events, that has to be present when the cause is effective. There is often a rubric for this; for example, in criminal (and most other) investigations into human agency, we use the rubric of means/motives/opportunity to get from the motives to the list of suspects. The list of modus operandi is the magnifying lens that fleshes out the candidate causes from the LOPC so that we can start fitting them to the case or rejecting them, for which we use the next premise.
- The third premise comprises the 'facts of the case,' and these are now assembled selectively, by looking for the presence or absence of factors listed in the modus operandi of each of the LOPCs. Only those causes are (eventually) left standing whose modus operandi are completely present. Ideally, there will be just one of these, but sometimes more than one, which are then co-causes.

What and When can Theory-Based Evaluation Contribute?

a) Before Implementation

To learn about the plausible effectiveness of a new intervention, an analysis of the theory underlying the intervention can be done. The evaluation tries to open the black box of the intervention: what are the mechanisms that are believed to make the intervention work? How plausible is it that these mechanisms ´do the job´? To detect these mechanisms, one has to search in documents, interviews, transcripts and speeches (of policy-makers, civil servants, etc.) for statements that answer the question why it is believed (or hoped) that the new intervention will make a difference.

It is crucial to be clear about what mechanisms are. Mechanisms are not the input-output-result process-variables, nor are they the dimensions usually contained in logical frameworks, logic models or statistical equations. Coleman (1990) and others point to three types of mechanisms: situational, action-formation mechanisms and transformational.

- Situational mechanisms operate at the macro-to-micro level. This type of mechanism shows how specific social situations or events help shape beliefs, desires, and opportunities of individual actors. An example is the opportunity structure a community, village or city - the more there are opportunities (for crime, for unemployed), the larger the chance that crimes will be carried out and jobs will be found.
- Action-formation mechanisms operate at the micro-to-micro level. This type of mechanism looks at how individual behavioral choices and actions are influenced by specific combination of desires, beliefs, and opportunities. Examples are cognitive biases (cognitive dissonance, fundamental attribution error), incentives (rational choice, exchange).
- Transformational mechanisms operate at the micro-to-macro level and show how a number of individuals, through their actions and interactions, generate macro-level outcomes. An example is 'cascading' by which people influence one another so much that people ignore their private knowledge and instead rely on the publicly stated judgments of others. The 'bandwagon phenomenon'— the tendency to do (or believe) things because many other people do (or believe) is related to this, as are 'group think' and 'herd behavior'.

One option to find information on mechanisms is to carefully read documentation and search for statements indicating the (espoused) motivations or rationales behind an intervention ('we believe that…', 'it is generally accepted that this option is …', 'based on our experience with the policy field, we decide that …'; 'the only real option to address this problem is ...'). One can apply content analysis to do this work. However, often mechanisms are not described clearly; they can only be found by reading between the lines and by applying argumentation analysis (Leeuw, 2003). Various argument ('assumption') visualization software applications can be used to detect arguments and order them and (logically) relate them to one another.

Once the mechanisms have been detected, the next step is to compare the statements, assumptions or beliefs (of policy-makers) about mechanisms with evidence from review and synthesis studies. Put differently: compare policy beliefs about mechanisms with research-based evidence. The evidence can be found in repositories like the Campbell Collaboration, the UK Evidence Network, the What Works Clearing House and others (see Hansen & Rieper, 2010 for an overview), but also (meta) search databases like the Web of Science are relevant.

The more the mechanisms believed by policy makers to be at work are in line with research-based evidence, the greater the plausibility of the new intervention to be effective.

Example 1: Subsidies

Pawson (2002) categorized six subsidies, covering incentives to stimulate fire alarm installation in homes, to give up smoking, to widen educational opportunities for students, to improve property, to help ex-offenders to re-socialise and to reduce inner city environmental pollution through subsidizing free-city-centre bikes. Next, he inventoried evaluations of these subsidies and studied the role of situational (context) mechanisms in understanding the success or failure of the subsidies. He produced a list of nine context factors that contribute to success or failure. In order to judge the plausibility that the new subsidy would be effective, Pawson’s context factors can be compared to the assumed context mechanisms of the new subsidy. The more the new subsidy takes into account the context mechanisms Pawson found in evaluations of (relatively) successful subsidies, the more plausible it is that the new subsidy will be effective.

Example 2: Fear-Arousal Communication and Behavior Change

This example deals with the fight against cocaine smuggling through people swallowing the drug and travelling between the Dutch Antilles and the Netherlands. Young deprived men were paid through organised crime several thousand Euros to fly between the Antilles and Amsterdam using the internal concealment method, i.e., swallowing small balls filled with cocaine and delivering the ‘stuff’ in Holland. The Dutch government was successful in reducing this kind of drug trafficking through an almost 100% control of passengers arriving at Amsterdam who came from certain regions. However, the policy was expensive, which made officials to think about an alternative. Could a public information campaign using leaflets, mass media and local media that present fear-arousal information about the medical dangers of the internal concealment method and the likelihood to be arrested, be an effective (and less expensive) intervention to reduce trafficking?

Kruisbergen (2007) evaluated this policy idea. He synthesised results from evaluations of the impact of ‘fear-arousal health education programmes’ in general (about smoking, dangerous drinking, etc.) and compared the mechanisms and contexts found in these studies with the existing empirical information about the contexts in the Dutch Antilles and some of the social and behavioral characteristics of ‘cocaine swallowers’. There was a huge discrepancy between contexts and mechanisms of successful fear-arousal communication health campaigns and the specific characteristics of cocaine swallowers and their contexts. Crucial conditions that made fear-arousal communication have an impact on (health) behavior did not exist in the case of cocaine swallowing behaviour. Kruisbergen's conclusion was that the likelihood of preventing illegal `immigration´ of cocaine to the Netherlands by implementing a public awareness campaign would be small to very small. In other words: the plausibility of the theory that fear-arousal communication will reduce drug trafficking using the internal concealment method was very limited

Example 3: Educational Governance

Janssens & de Wolf (2009) carried out an ex ante evaluation of the theory underlying a new Dutch educational policy that combines accountability and inspections. A central feature of this policy is that it strives for an optimal balance between accountability, inspections, self-evaluation and improvement activities. The programme is called 'educational governance'. It stimulates systems of internal quality assurance by (a) establishing national standards and public accountability, (b) encouraging parents to take an active role in internal supervision processes within schools (through boards of trustees), and (c) enacting external government supervision. One of the objectives is to make schools take a proactive role in educational accountability as opposed to a reactive one. Another objective is to involve other actors (parents, students and teachers) in the accountability system.

The authors applied approaches developed by Pawson and Tilley (1997), Weiss (2000), Leeuw (2003) and others. The evaluation consists of three parts. First, it reconstructs the theory that underlies the aims and policy of educational governance. It identifies the central assumptions of the policy and uses them to reconstruct its causal scheme, 'reconstruction of the programme theory'. The second step is an assessment of the main assumptions of the programme. This involves assessing the tenability of these assumptions in light of the most recent research based insights. With this, the evaluation ascertains the acceptability and empirical tenability of the ideas or assumptions and the validity of the logic underlying the programme theory. The more 'suitable' and 'evidence-based' the assumptions are, the greater the chance that the programme theory will work in practice. In the last step, the evaluation combines and weighs the conclusions of the evaluations of the separate assumptions. It also determines the mutual compatibility of the assumptions. This last step explores if the programme will be able to generate the intended effects. It also helps to identify theoretical imperfections or other threats to the effectiveness of the programme in practice.

The evaluation found that the policy might not achieve its objectives and identified the elements which needed improvement. A flaw in the theory underlying the programme was found, which threatened its potential effectiveness. Furthermore, the evaluation showed that there was a risk of contrary and incompatible interests among actors, as well as some practical reasons why the programme might not work.

Conclusion

To answer the plausibility question ex ante, it is suggested to focus on mechanisms as the 'drivers' of the new policy or intervention and then compare these with already available research and evaluation evidence for these or similar mechanisms. The more the intervention theory is backed by evidence on working mechanisms, the more plausible the theory is and the likelier it is that the new policy will make a difference.

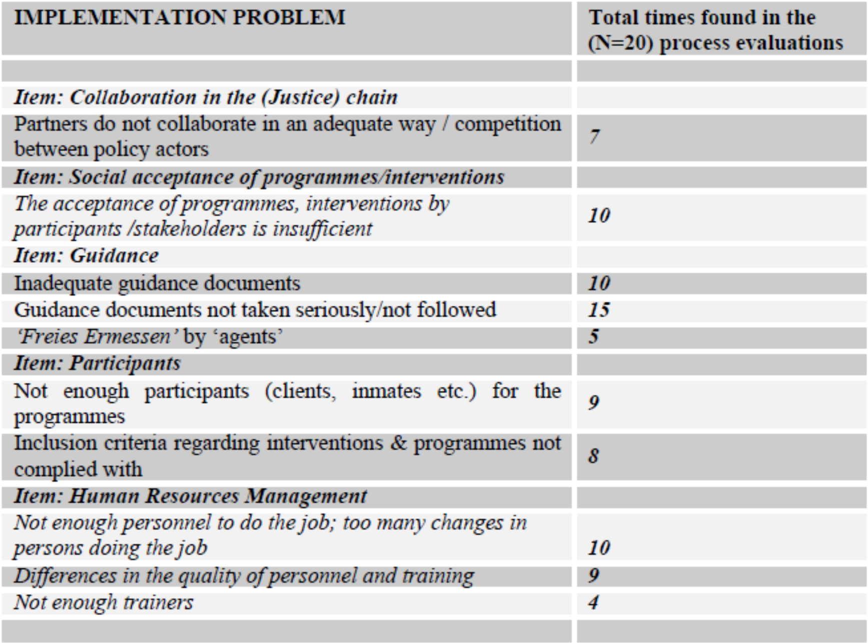
b) During Implementation

What can TBE contribute to find out - during implementation - how plausible it is that a policy or programme will be effective? Two routes can be distinguished.

The first route focuses on the implementation theory, i.e., the theory that describes which operations have to be performed and which (organisational) conditions have to be met for a new intervention be ´put to work´. There is abundant evidence that when 'programme integrity' is limited, which means that the implementation of the policy is not as was planned, this reduces the effectiveness of the policy (Barnoski, 2004; Carroll, Patterson, Wood, Booth, Rick & Balain, 2007; Nas, van Ooijen & Wieman, 2011). Take the example of a new intervention on e-learning. Such an intervention not only needs to be based on sound ('working') mechanisms, but several practical problems also have to be solved. The ICT infrastructure, including bandwidth has to be available and ready; staff, teachers and parents have to accept the new approach; software programs have to be available; students have to work with them and side-effects have to be understood. An example of a side effect is the time it can take to train (older) staff members to become familiar with e-learning and to be able to coach the educational processes.

In a recent Dutch meta-study of 20 implementation evaluations of interventions in the world of crime and justice, the following implementation problems were found to happen most often (Nas, van Ooyen & Wieman, 2011; Leeuw, 2011).

Table 2 Incidence of problems when implementing (penal) sanctions/ behaviour (modification) programmes (based on 20 Dutch process evaluations).



IMPLEMENTATION PROBLEM	Total times found in the (N=20) process evaluations
Item: Collaboration in the (Justice) chain	
Partners do not collaborate in an adequate way / competition between policy actors	*7*
Item: Social acceptance of programmes/interventions	
The acceptance of programmes, interventions by participants /stakeholders is insufficient	***10***
Item: Guidance	
Inadequate guidance documents	***10***
Guidance documents not taken seriously/not followed	***15***
'Freies Ermessen' by 'agents'	***5***
Item: Participants	
Not enough participants (clients, inmates etc.) for the programmes	***9***
Inclusion criteria regarding interventions & programmes not complied with	***8***
Item: Human Resources Management	
Not enough personnel to do the job; too many changes in persons doing the job	***10***
Differences in the quality of personnel and training	***9***
Not enough trainers	***4***

The implementation of these interventions did not take into account the likelihood that factors such as social acceptance, lack of guidance, collaboration problems and personnel problems would cause problems. The more the theory underlying implementation takes account of these and similar implementation problems, and how to prevent or reduce them, the greater the likelihood of the new programme being successful (if, of course, the intervention theory is plausible). When no information is available on the problems, the plausibility of the intervention being effective is reduced.

The second route for TBE to assist in evaluations during implementation is described by Kautto & Simila (2005) and focuses on the intervention theory and 'recently introduced policy instruments' (RIPI’s). When evaluators are confronted with the request to assess the (future) impact of policies, this is understandably difficult. It takes time for an intervention to be fully implemented and ‘working’. As evaluation time is not similar to political time, this poses problems for policy-makers, evaluation commissioners and evaluators. Kautto and Simila (2005) presented an approach in which a central role is given to the intervention theory. They evaluate a change of the Environmental Protection Act (1999) in Finland. A European Union Directive on Integrated Pollution Prevention and Control was transposed into the Finnish legal system. At the core of the reform was the integration of five different permits (air pollution, water pollution, waste management, protection of health and neighbourhood relations) into one environmental permit. To establish the (final) impact on the environment of the new legal arrangement including the reduction of permits would take years. Waiting that long was not an option, so the evaluators did something else. They started an evaluation relatively soon after the announcement of the new Act. They unpacked the intervention theory (‘why will a reduction from five permits to one be effective in terms of environmental protection?’), distinguished between outputs and results of the Act and collected data on outputs that were already available. Information on results was not yet available. They checked the plausibility of the part of the intervention theory that linked certain characteristics of permits to the final goal of the new Act (environmental impact/outcomes). Because more than 600 permits (= outputs) had been granted during the first two years of implementation, it was possible to assess whether the assumptions about the characteristics of the outputs were correct. Kautto and Simila: 'this enabled us to say something important about the effectiveness despite the fact that the (final) results had not yet occurred. Concurrently, it must be noted that while the permits have not been changed as assumed at the beginning of the implementation process, this does not mean that they will never be changed. The evaluation itself may have an impact on the implementation and as a result, or for other reasons, the authorities may place greater emphasis on gaps and priorities in the future. In this context, the intervention theory was not used to predict the future, but to guide the evaluation'.

An interesting conclusion Kautto and Simila draw is that although an impact analysis was not possible because results had not yet occurred, this does not necessarily imply that the use of impact as a criterion is also impossible – thanks to the concept of an intervention theory. However, the content of the effectiveness criterion must be reformulated. As the evaluators have shown, effectiveness refers to the degree of correspondence between intended policy goals and achieved results. If results have not yet occurred, a comparison of the objectives and achieved results is impossible. But what is possible, instead, is to ask whether the outputs include features that are preconditions to the achievement of the goals according to the intervention theory. The more this is the case, the greater the probability that effectiveness is in reach.

Conclusions

To answer the impact question during the implementation process can be done in two ways. The first is to study the implementation theory and check to what extent this theory takes account of pitfalls that can often be found when implementing interventions. The second route is to follow an approach articulated by Kautto and Simila (2005) known as RIPI-evaluations (‘recently implemented policy instruments’).

c) After Implementation

What can TBE contribute ex post to establishing the counterfactual, when it is not feasible to use experimental or non-experimental designs?

If an experimental setting is not possible, if natural experiments or different designs of non-experiments are not possible, then one can move to more qualitative approaches to establish a counterfactual.

- Use the counterfactual history approach and hypothetical question-studies

Counterfactual history, also sometimes referred to as virtual history, attempts to answer 'what if' questions. It seeks to explore history and historical incidents by means of extrapolating a timeline in which certain key historical events did not happen or had an outcome which was different from that which did in fact occur. Fogel (1964) looked at where America would have been (in terms of its GDP) had there been no railroads. He hypothesised that the increase in GDP, given by the railroads, would have happened anyway had other technologies taken hold. Examining transportation costs for primary and secondary goods, he compared the 1890 economy to a hypothetical 1890 economy in which transportation infrastructure was limited to wagons, canals and rivers. Fogel found that the impact of the railroads was small - about 7% of the 1890 GDP. A substitute technology, the more extensive canal system, would have been able to reach a comparable economic growth. After Fogel many other counterfactual historical studies were published, including work that combines experimental psychology and history.

Methodological rules of thumb are available on how to do this work and how to judge its quality (Tetlock & Belkin, 1996). These authors also collect data from hundreds of experts that predict the counterfactual future/past. By analysing their answers, patterns and (ultimately, when time progresses) the validity of their statements, these researchers are trying to unpack what did the 'work' in predicting the future (and vice-versa: the past).

For evaluators, a similar approach is possible. If, for example, the impact of a grant to companies to stimulate innovation or a new system of knowledge brokers for SME has to be assessed and statistical evaluation designs are not possible, evaluators can be asked to develop a counterfactual for the situation had there been no grants. Answering the question can be done in line with the way historians work (using existing data and theories), but it is also possible to apply the hypothetical question-methodology, known from policy acceptance studies and marketing. The question then is what people would do if policy \`a\` or \`b´, was not implemented. An early example is the Thompson & Appelbaum (1974) study of the impact of population policies in the USA, which was later one of the pillars upon which Dutch hypothetical question evaluations of population policies were built (Moors et al, 1985).

- Apply contribution analysis

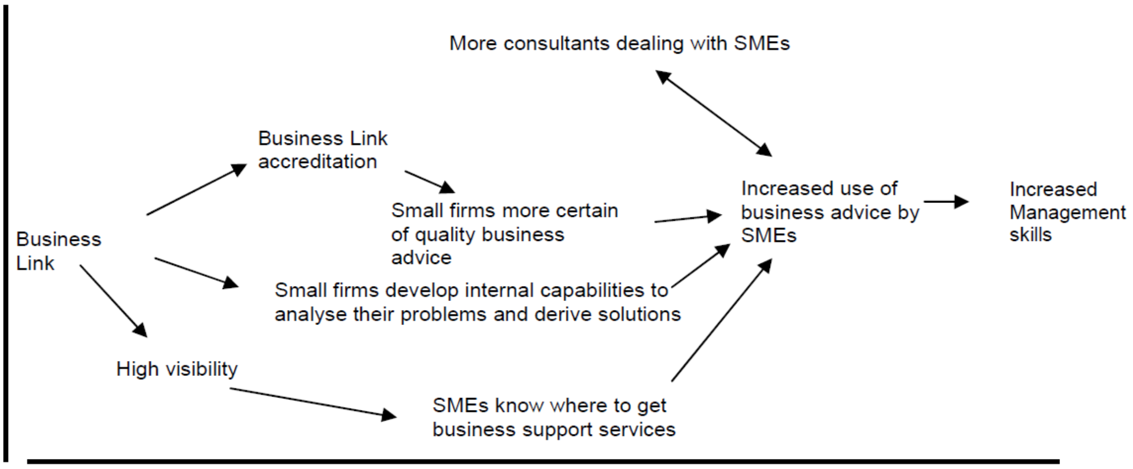
Contribution analysis is based on the existence of, or more usually, the development of a theory of change for the intervention being examined. A theory of change sets out why it is believed that the intervention's activities will lead to a contribution to the intended results; that is, why and how the observed results can be attributed to the intervention. The analysis tests this theory against logic and the evidence available on the results observed and the various assumptions behind the theory of change, and examines other influencing factors. It either confirms the postulated theory of change or suggests revisions in the theory where the reality appears otherwise. It is best done iteratively, building up over time a more robust contribution story. The aim is to reduce uncertainty about the contribution the intervention makes to observed results through an increased understanding of why the results have occurred (or not) and the roles played by the intervention and other factors. Mayne has outlined steps of contribution analysis including examples in different fields (see section 2.3).

Although the authors did not relate their study to contribution analysis, Mole's et al (2007) large-scale telephone survey of over 3,000 SMEs and 40 face-to-face interviews with business owners, which tried to assess the impact of Business Link Organization (BLO) activities on businesses that received assistance, is somewhat linked to this approach.

Evaluating the Impact of Businesses Link Organization activities (Mole et al, 2007)

BLO is a type of small business support activities that some (European) governments have implemented. It can be seen as a type of brokerage and is usually targeted at small (to medium size) businesses. BLO activities are believed to increase economic productivity and job growth. The small business service aims to build the capability for small business growth and the advice and support provided by Business Links are intended to improve the management skills and thereby improve business performance and entrepreneurship. The evaluation by Mole et al (2007) paid some attention to the programme theory underlying the BLO (see figure below). In the figure the arrows indicate a direction of causality (Mole et al, 2007: 28).

Figure 1.1: Programme Theory for Business Links


More consultants dealing with SMEs
Business Link accreditation
Small firms more certain of quality business advice
Increased use of business advice by SMEs
Increased Management skills
Business Link
Small firms develop internal capabilities to analyse their problems and derive solutions
High visibility
SMEs know where to get business support services

Although the evaluators refer to this programme theory and present several hypotheses on relationships between BLO and dependent variables like increased management skills and the possibility of finding business advice, no attention was paid to the question how and to what extent the programme (theory) is empirically linked to policy goals like productivity increase or job creation. Had these topics been added to the (large scale) empirical approach of the evaluation, it would have made the relevance of the study larger[12].

- Work in line with expert judgments

Expert judgments or connoisseur evaluations are used to cover strategies that pool the opinions of experts to assess performance of programmes or policies. Recent forms of expert judgment include:

[12] Also, a number of other (methodological) problems in the study would have to be addressed, like the quality of data (collection).

- Accreditation and evaluation activities of the effectiveness of behavioural modification programmes or educational programmes. This is a combination of expert judgment and meta-evaluation & meta-analysis of what is known about – for example - programmes reducing violence in public places by people.
- Civic evaluation, based on ‘the wisdom of the crowd’ that evaluates policies and organisations. Here the experts are ‘the people’ those engaged in these social groups.

If one wants to use this approach to develop the counterfactual, it must be stressed that the evaluators coordinating this effort are not looking for answers from experts on the question what the impact of intervention X would have been, but exactly on the opposite question: what would have happened without intervention X.

- Work in line with the GEM: 'general elimination methodology' when using one of these approaches.

GEM was coined by Scriven (1978; 2008) – see section 2.7. If an evaluation has found results on impact although the design is weak or insufficient, and if the underlying intervention theory is relatively plausible, the GEM can be used to check if there are other factors (than the intervention) that are more plausible as explanations of the impact. The primary goal of GEM is to see how solid the arguments are, indicating that the intervention caused or contributed to the outcomes. A GEM evaluator invites the ‘believers’ in the contribution of the intervention to discuss alternative explanations, having nothing to do with the intervention. The more believers serious challenge and falsify these alternative explanations, the more plausible it is that the intervention indeed is causing the difference. Simultaneously the GEM evaluator tries to falsify the intervention theory. The more successful they are in doing that, the less plausible this theory is.

What can TBE contribute ex post when an impact evaluation, including the counterfactual established through experimental or non-experimental designs, has been carried out, but an explanation of the findings is lacking?

Evaluators applying experimental or non-experimental designs do not always pay attention to the social and behavioural mechanisms that underlie the interventions they assess. The interventions, Pawson and Tilley (1994) claim, are seen almost as black boxes, whereas to understand why things work (assuming they do), one needs to know which social and behavioural mechanisms are active and in which contexts (Pawson & Tilley, 1997). During the 1990s and early 2000s an almost paradigmatic conflict existed between (some) experimentalists and (some) realists on this topic. More recently this conflict has become less severe.

What can be done to remedy the lack of explanations? The first answer is to open up the black boxes afterwards in the way suggested above. The second answer is to combine experimental and non-experimental impact evaluations and realist evaluations. An example of this approach can be found in a paper by Van der Knaap et al (2008). It describes an approach that combines the use of the 'Campbell collaboration standards' with the realist notion of addressing contexts-mechanisms-outcomes (CMO) that underlie interventions (see section on realist evaluation).

The Campbell Collaboration (C2) is an international volunteer network of policymakers, researchers, practitioners, and consumers who prepare, maintain and disseminate systematic reviews of studies of interventions in the social and behavioral sciences (see http://www.campbellcollaboration.org). The organisation is named after Donald T. Campbell, an

American social scientist and champion of public and professional decision-making based on sound evidence. C2 reviews are designed to generate high-quality evidence in the interest of providing useful information to policy-makers, practitioners and the public on what interventions help, harm or have no detectable effect. The organisation has developed standards for systematic review (clear inclusion/exclusion criteria, an explicit search strategy, systematic coding and analysis of included studies, meta-analysis).

Van der Knaap et al focused on interventions to prevent or reduce violence in the public domain. To merge 'Campbell standards' and the realist evaluation approach, the realist approach was applied after finishing the Campbell-style systematic review. The following box describes the way the ‘merger’ works:

> 'Our first goal was to provide an international overview of effective or at least promising measures to prevent violence in the public and semi-public domains. The second goal was to gain insights into the behavioural and social mechanisms that underlie effective or promising prevention measures and the circumstances in which these are found to be effective. We defined violence as “the deliberate use of physical strength or power and/or the threat thereof, aimed against another person or group of persons and which results or is likely to result in injury, death or psychological damage"' (van der Knaap et al., 2006, p. 21).
>
> 'Our first step was to conduct a Campbell-style review. …….We collected 48 studies that met our inclusion and exclusion criteria. These 48 publications relate to 36 interventions, most of which are designed to prevent violence in schools. We did not include a meta-analysis in our study but instead assessed each study’s methodological quality using the Maryland Scientific Methods Scale (MSMS). Only experimental (level 5), quasi-experimental (level 4), comparative designs without matching or randomisation (Level 3) and evaluations using a before-after design (level 2) were used. 13 Level 2 studies, 10 Level 3 studies, 13 Level 4 studies, and 11 Level 5 studies therefore were included. For information on the MSMS, see http://www.ncjrs.gov/pdffiles/171676.PDF).
>
> Based on the MSMS scores, we classified each of the 36 interventions into one of the categories of effective, potentially effective, potentially ineffective, and ineffective. However, not all studies could be grouped into one of the four categories. In 16 cases, the quality of the study design was not good enough to decide on the effectiveness of a measure. Nine interventions were labelled effective and six were labelled potentially effective. Four interventions were labelled potentially ineffective and one was labelled ineffective in preventing violence.
>
> After finishing our Campbell-style review, we applied the realist approach to each of the interventions in our study. This proved to be rather difficult, for a lot of information was missing in the original publications. Often, no explicit theory was described underpinning the intervention, and information on mechanisms and context was scarce. By having two researchers read the publications and identify implicit and explicit statements pertaining to mechanisms and context, we tried to reconstruct CMO configurations. Among other strategies, we scrutinized the outcome measures that were used by the evaluators. For instance, if they focused on attitudes and knowledge, we argued that the program designers meant to achieve changes in attitudes and knowledge and assumed that these changes would cause behavioural change. Whereas some publications offered more detailed information, the mechanisms we identified could mostly be described in general terms only. Based on the evaluations we analysed, some ten mechanisms could be identified that, in fact, boiled down to the following three:
>
> The first of these is of a cognitive nature, focusing on learning, teaching, and training. The

second overarching mechanism concerns the way in which the (social) environment is rewarding or punishing behaviour (through bonding, community development, and the targeting of police activities). The third mechanism - of a more general nature - is risk reduction, for instance, by promoting protective factors' (Van der Knaap et al, 2008:55).

Van der Knaap et al (2008) summarise their view on the practical importance of their work as follows: 'Combining the approach outlined by the Campbell Collaboration and the realist evaluation approach is commendable in several ways. First, the result of applying Campbell standards helps to distinguish different types and (methodological) levels of evaluation designs. For those interested in the impact or effectiveness of interventions, this is important. Second, the opening up of the micro-architecture of those interventions that have been shown to be effective, or at least potentially effective, helps better understand what makes these interventions work. Moreover, by also studying the mechanism-context configurations of interventions that appear to be ineffective, one can learn more about the conditions that are necessary for mechanisms to work. A third advantage of our combination, which we have ourselves not been able to realize, is that by applying a realist synthesis approach (Pawson, 2006), knowledge from outside the field of crime and justice evaluations but of direct relevance to the mechanisms can be used to understand why (some) programs work and others do not. In the longer run, these combinations of knowledge funds will help in understanding the interventions better and will probably also help in designing better interventions'.

Conclusions

When an experimental or non-experimental counterfactual is not feasible, TBE can help to derive a counterfactual in several ways. One is to apply the approach of counterfactual historians; another to use hypothetical-question studies while a third is to involve expert judgements ('connoisseur evaluations'). The General Elimination Methodology is also recommended.

When an experimental or non-experimental impact study has been done and results on the effectiveness of the policy or programme are available, attention is not always paid to the why and the how question. These questions are important for policy makers. TBE can help to find explanations: first, by opening up the black boxes of the interventions evaluated and searching for working mechanisms; second, by doing a (follow-up) study in which evaluation designs working with experimental or non-experimental counterfactuals are combined with the realist evaluation approach from the start.

Indicators

What can TBE contribute to define and operationalize performance indicators of policies or programmes?

TBE and performance indicators are related, but it is a rather complex relationship. There are at least three sets of theories involved.

The first is the theory that the improvement of performance of organisations is stimulated by working with indicators. Mechanisms are that indicators drive workers and management to work (more) in line with the goals set by the organisation and in such a way that comparisons between divisions, departments and outside organisations are possible. It is also believed that indicators stimulate 'learning'.

A second theory specifies the form, content and number of indicators and points to intended and unintended effects of working with them. Some indicators may be better in contributing to learning, while others may stimulate bureaucratization, red tape, dramaturgical compliance and the performance paradox (van Thiel & Leeuw, 2002).

We do not deal with these two sets of theory here. Instead we discuss the question of what policy/programme theories can contribute to developing and implementing effective indicators.

As working with indicators has become mainstream in policy fields and as evidence is available on 'pathologies' that go hand in hand with using indicators (Bouckaert & Balk 1991; van Thiel & Leeuw, 2002), the role policy/programme theory plays in the world of indicators is important.

Fifteen years ago Bickman (1996) suggested that a logical starting point for developing the most appropriate indicators was to create a model or programme theory. Birleson, Brann & Smith (2009) did exactly that in a paper on clinical and community care in Child and Adolescent Mental Health Services (CAMHS) in hospitals. They articulated the programme theory of different services by looking at programme operations, proximal outcomes and final outcomes and relationships between them. They showed that without an articulated programme theory, indicators were likely to be less relevant, over inclusive or poorly linked with the programme operations they aimed to measure. In a rather different field (road safety performance) Hakkert, Gitelman & Vis (2007) did something similar. This study provides details about the theory behind the development of Safety Performance Indicators in seven major areas which are central to the fields of activity in road safety in Europe.

A third example is different as it shows how problematic the use of indicators can be if there is discrepancy between the programme theory and the indicators. Lindgren (2001) shows how thin the ice is for performance measurement, when indicators are developed and used without taking notice of the (richness of the) programme theory. The case concerns popular adult education in Sweden and demonstrates that important activities and characteristics of adult education are not covered by the key indicators, which leads to pitfalls in the use of the performance data, while there is also a set of indicators not linked to any substantive part of the programme theory.

In linking TBE with performance indicators, four activities are central:

- The first is to (re)construct the theory underlying the policy or programme and to develop indicators that cover the richness of the underlying theory.
- The second activity is to understand that indicators can trigger behavioural responses that can lead to a 'performance paradox': organisations good in measuring performance indicators are not necessarily the most effective organisations. Examples of these trigger mechanisms are the following:
    - Emphasising - by policy-makers/principal - that compliance with protocols and procedures is crucial (often leading to the production of data largely to satisfy the principals' need for sound protocols and procedures);
    - Having to work with elusive and contradictory policy goals;
    - Having to work with goals that are inherently not or very difficult to operationalise and measure.
- These trigger mechanisms can contribute to an unintended performance paradox. Van Thiel & Leeuw (2002) also point to the problem that there are mechanisms leading to an intended performance paradox. These are 'cognitive sabotage' of performance measurements and audits, including 'cooking' the data, 'creaming' (focussing on the best cases) and myopia (only information on short term objectives is presented, while more information is available).

- The final step is to prevent the performance paradox. Meyer and Gupta (1994) recommend the use of targets and comparisons over time, between organisations or between different units within the same organisation.

Conclusions

The more important performance indicators are, and the more there is evidence that working with them can have unintended and undesirable side effects, the more it is relevant that TBE is used when designing and implementing them. If not, the likelihood that indicators are distanced from the operations and mechanisms of the policies and programmes analysed will increase.

These conclusions apply directly to EU Cohesion policy programmes. With a growing demand for a more performance-oriented EU Cohesion policy, the importance of performance indicators also increases. This requires a greater focus to be put on indicators which should reflect the objectives of the policy and better capture the effects of the interventions. This new approach is promoted in the context of the 2014-2020 programming exercise (for more information, please refer to: http://ec.europa.eu/regional_policy/impact/evaluation/performance_en.cfm).

Problems to be avoided when doing TBE

The following pitfalls or problems when doing TBE can be mentioned. If evaluators are not aware of them, TBE will create 'error costs'[13].

- Avoid sloppy reconstructions and tests of underlying programme theories.

Tilley (1992) brought together several practices that contribute to the production of error costs when doing evaluations Sloppy reconstructions and tests of underlying programme theories (`misconstrue programmes') is one; neglecting contextual differences when comparing results from evaluations (in different time periods) is another. Misinterpretation of what caused a programme not to work (by confusing implementation problems, measurement problems and difficulties with the programme theory) is another (Tilley, 1992). Programmes that could have been effective are sometimes terminated or considered not ready for implementation because of a faulty theory-reconstruction. Error costs involved are inefficiency, foregone investments in developing the programme and wasted money on behalf of the evaluation, while an opportunity cost is that the social problem to be remedied by the programme, continues to exist. Related to this is what Funnel and Rogers (2011) call the 'No Actual Theory' trap: an evaluator refers to programme theories which are in fact not theories at all. '[Instead], they simply display boxes of activities and boxes of results without demonstrating logical and defensible relationships between them and the various items listed in the boxes'.

- Take notice of the problem of concatenation of mechanisms and try to solve it.

Hedstrom (2005) has dealt with this point: 'it is often necessary [when doing a TBE] to consider several mechanisms simultaneously in order to make sense of a specific social phenomenon'. He adds that 'these mechanisms may interact with another in a complex way' (ibid). In recent work by Rogers (2009) and Rogers & Funnel (2011), attention is paid to the relationship between complexity, complicatedness and theory-based evaluations.

- Prevent 'designed blindness'.

This happens when practitioners and evaluators are focused on the programme theory in such an intense way that they not only start to frame every activity of the evaluated programme in terms of this theory (Friedman, 2001), but also start to believe that the intervention theory is inherently

[13] See Leeuw (2010) in the 'Zeitschrift für Evaluation' on costs and benefits of evaluations.

'valid' and 'good'; this point is related to the psychological mechanisms of tunnel vision. The error cost is that the evaluation ends up being circular: as the evaluator and the evaluated programme are 'captured´ in the programme theory, the possibility for a serious test of the theory by collecting data for example, is very small.

- Prevent the 'polishing' up or quasi-enrichment of the policy theory.

This happens when policy-makers ask evaluators to polish up or 'enrich' assumptions underlying their policies, while in reality the policies are grounded on rather thin assumptions. The error costs are twofold: first, it resembles impression management (the 'rich' and informative intervention theory forms the fundament of an intervention that is largely a 'show policy') and, secondly, it can set in motion a process of imitation in organisations that will create future failures and faulty processes.

- Not using the programme theory for evaluation.

Funnel and Rogers (2011) refer to this 'trap'. It concerns the discrepancy between developing or reconstructing the programme theory but nevertheless doing the (empirical) evaluation without paying attention to this theory. It can be labelled a case of 'wasted words'.

What Theory Based Evaluation is not?

TBE is not the same as presenting:

- A logical framework;
- 'Unexplained causal arrows' and
- Schemes such as the input-throughput-output-diagram often used in (performance) measurements and auditing (Astbury & Leeuw, 2010).

Selected References

Astbury, Brat & Frans L Leeuw (2010). Unpacking black boxes: mechanisms and theory-building in evaluation. American Journal of Evaluation 31 (3): 363-381.

Barnoski, R. (2004). Outcome evaluation of Washington State's research-based programs for juvenile offenders. Washington State Institute for Public Policy.

Birleson, P., Brann, P. & Smith, A. (2001). Using program theory to develop key performance indicators for child and adolescent mental health services. Australian Health Review, 24 (1): 10-21.

Bouckaert, G.& Balk,W. (1991). Public productivity measurement: Diseases and cures. Public Productivity & Management Review, 15 (2): 229-235.

Carroll, C., Patterson, M., Wood, S., Booth, A., Rick, J., & Balain, S. (2007). A conceptual framework for implementation fidelity. Implementation Science, 2(40). http://implementationscience.com/content/pdf/1748-5908-2-40.pdf

Carvalho, S. & White, H. (2004). Theory-based evaluation: the case of social funds. American Journal of Evaluation, 25, (2): 141-160.

Chen, H. T., & Rossi, P. H. (1980). The multi-goal, theory-driven approach to evaluation: A model linking basic and applied social science. Social Forces, 59, 106-122.

Coleman, James (1990). Foundations of Social Theory, Belknap Press.

Donaldson, S. I. (2007). Program theory-driven evaluation science. New York, NY: Lawrence.

Fogel, R. (1964). Railroads and American Economic Growth: Essays in Econometric History (1964).

Hakkert, A.S, Gitelman, V. and Vis, M.A. (Eds.) (2007). Road Safety Performance Indicators: Theory. Deliverable D3.6 of the EU FP6 project SafetyNet.

Friedman, Victor, Designed Blindness: An Action Science Perspective on Program Theory Evaluation. American Journal of Evaluation, 22:161–181.

Hansen, H. & O. Rieper (2010). Institutionalization of Second-Order Evidence-Producing Organizations, in O. Rieper et al. (eds), The Evidence Book: Concepts, Generation and the Use of Evidence, pp. 27-52.

Janssens, Frans & Inge de Wolf (2009).Analysing the Assumptions of a Policy Program. An Ex-ante Evaluation of ''Educational Governance'' in the Netherlands. American Journal of Evaluation 30(3): 411-425.

Kautto, P. and Similä, J. (2005). Recently Introduced Policy Instruments and Intervention Theories. Evaluation, 11(1): 55–68.

Kruisbergen, E.W. (2005). Voorlichting: doen of laten? Theorie van afschrikwekkende voorlichtingscampagnes toegepast op de casus van bolletjesslikkers. Beleidswetenschap, 19, 3.

Leeuw, Frans L. Policy theories, knowledge utilization, and evaluation. Knowledge and Policy, 4: 73-92.

Leeuw, Frans L. & J.E. Furubo (2008). Evaluation systems: what are they and why study them? Evaluation, 14 (1): 157-169.

Leeuw, Frans, (2009). Evaluation Policy in the Netherlands. New Directions for Evaluation, 123: 87-103.

Leeuw, Frans L. (2003). Reconstructing program theories: methods available and problems to be solved. American Journal of Evaluation, 24 (1): 5-20.

Leeuw, Frans L. (2011). Can legal research benefit from evaluation studies? Utrecht Law Review, 7 (1): 52-65.

Leeuw, Frans & J. Vaessen (2009). Impact evaluation and development. NONIE & World Bank, Washington.

Lindgren, L. (2001). The Non-profit Sector Meets the Performance-management Movement; A Programme-theory Approach. Evaluation, 7(3): 285–303.

Ludwig, J. Kling, J.R, and Mullainathan, S, (2011). Mechanism Experiments and Policy Evaluations. Journal of Economic Perspectives, 25 (3): 17–38.

Mole, Kevin et al, Economic Impact Study of Business Link Local Service, University of Warwick, 2007.

Nas, Coralijn, Marianne M.J. van Ooyen-Houben & Jenske Wieman (2011). Interventies in uitvoering. Wat er mis kan gaan bij de uitvoering van justitiële (gedrags)interventies en hoe dat komt. WODC Memorandum, Den Haag.

Patton, Michael (2008). Advocacy Impact Evaluation. Journal of multidisciplinary Evaluation, 5 (9): 1-10.

Pawson, Ray (2002), Evidence-based Policy: The Promise of `Realist Synthesis´. Evaluation 8 (3): 340–358

Pawson, Ray (2003). Nothing as practical as a good theory. Evaluation 9(4): 471 – 90.

Pawson, Ray (2006 a). Evidence-based policy: a realist perspective. London.

Pawson, Ray (2006 b). Simple principles for the evaluation of complex programmes,' in A Killoran and A Kelly (eds), Evidence based public health . Oxford: Oxford University Press.

Pawson, Ray & Nick Tilley (1997). Realistic Evaluation, London.

Pawson, Ray & S. Sridharan (2010). Theory-driven evaluation of public health programmes, in: Evidence-based Public Health Effectiveness and efficiency, Edited by Amanda Killoran and Mike Kelly, chapter 4: 42-62.

Rogers, Patricia and Sue Funnell (2011). Purposeful Program Theory: Effective Use of Theories of Change and Logic Models. Jossey Bass.

Rogers, P. J. (2008). Using programme theory to evaluate complicated and complex aspects of interventions. Evaluation, 14, 29-48.

Rozendal, P., H. Moors & F. Leeuw (1985). Het bevolkingsvraagstuk in de jaren 80; opvattingen over overheidsbeleid, Nidi, Den Haag.

Scriven, M. (2008). Summative Evaluation of RCT Methodology: An Alternative Approach

to Causal Research. Journal of Multidisciplinary Evaluation 5(9), 11–24.

Scriven, M. (1976). Maximizing the Power of Causal Investigations: The Modus Operandi

Method, in: G. V. Glass (ed.) Evaluation Studies Review Annual, Vol. 1, Sage Publications, Beverly Hills, CA.

Suchman, E. (1967). Evaluative research. New York, NY: Russell Sage Foundation.

Tetlock, Philip E. and Aaron Belkin (1996). "Counterfactual thought Experiments in Global Politics: Logical, Methodological, and Psychological Perspectives." In Tetlock and Belkin (eds) Counterfactual Reasoning, Counterfactual thought experiments in global politics: Logical, Methodological and Psychological Perspectives. Princeton University Press, pp. 3-38.

Thiel, Sandra van & Leeuw, Frans L. (2002). The performance paradox in the public sector. Public Productivity and Management Review, 25: 267-281.

Thompson, V.D. & Appelbaum, M. (1974). Population Policy Acceptance: Psychological Determinants. Chapel Hill, N.C.: Carolina Population Center Monograph Series.

Tilley, Nick (1999). Evaluation and evidence-(mis)led policy. Evaluation Journal of Australasia, 11: 48-63.

US GAO (1995). Prospective evaluations methods, Washington.

US GAO (1986). Teenage pregnancy. 500,000 Births a year but Few Tested Programs. Washington.

Van der Knaap, Leontien M., Frans L. Leeuw, Stefan Bogaerts and Laura T. J. Nijssen (2008), Combining Campbell Standards and the Realist Evaluation Approach: The Best of two world. American Journal of Evaluation, 29 (1): 48-57.

Weiss, C. H. (1995). Nothing as practical as good theory: Exploring theory-based evaluation for comprehensive community initiatives for children and families. In J. Connell, A. Kubisch, L. B. Schorr, & C. H. Weiss (Eds.), New approaches to evaluating community.

Counterfactual Impact Evaluation

The Logic of Counterfactual Impact Evaluation

The Introduction to Impact Evaluation identified two separate sets of questions, one dealing primarily with programme and implementation theories and the second with quantification of effects. The first relying on theory-based methods, the second on counterfactual methods. In this section we deal exclusively with the second set of methods, devoted to *quantifying whether a given intervention produces the desired effects on some pre-established dimension of interest*.

Questions related to the sign and magnitude of programme impacts arise frequently in the evaluation of socio-economic development programmes. Do R&D subsidies increase the level of R&D expenditure by subsidized firms? Do targeted ERDF funds increase per capita income of the assisted areas? Do urban renewal programmes contribute to the economic development of urban neighbourhoods? Does support to SMEs increase their employment levels? Does investment in new public infrastructure increase housing values?

In other words, the evaluation problem has to do with the “attribution” of the change observed to the intervention that has been implemented. Is the change due to the policy or would it have occurred anyway? Answering these questions is not as straightforward as it might seem. The challenge for quantifying effect is finding a credible approximation to what would have occurred in *the absence* of the intervention, and to compare it with what actually happened. The difference is the estimated effect, or impact, of the intervention, on the particular outcome of interest (be it per capita GDP, R&D expenditure, housing values or employment levels).

Effects, impacts, and counterfactuals

A notation on terminology is necessary. Unlike in other evaluation settings, here impacts and effects are perfect synonyms. There is truly no meaningful difference between the two terms, they both refer to the notion of “causal effect”, the difference between the outcome occurred *after* an intervention has taken place and the outcome that would have occurred *in the absence* of the intervention. The popular distinction between "effects" as immediate results and "impacts" as long-run, or wider, effects, has no meaning on this context.

The counterfactual situation is purely hypothetical, thus can never be directly observed. For the same reason, an effect can never be directly observed, nor can an impact (impact indicators notwithstanding). By contrast, effects and impacts can be *inferred*, as long as the available data allows a credible way to approximate the counterfactual.

There are two basic ways to approximate the counterfactual: (i) using the outcome observed for *non*-beneficiaries; or (ii) using the outcome observed for beneficiaries *before* they are exposed to the intervention. However, caution must be used in *interpreting* these differences as the "effect" of the intervention.

Extreme caution is needed to interpret the observed differences as "effects"

These observed differences (over time, across individuals) indeed show "objective facts": for example, the performance of the supported firms *is* superior to that of the non-supported firms; the capital stock *has* increased after the support. What is problematic is the *interpretation* of these differences, what is dubious is their *causal interpretation.* Such interpretation *is* crucial for decision makers: only differences that have a plausible causal interpretation reveal "what works". For example, how much of the difference in outcomes between supported and non supported companies is due to the support received? And how much of the difference is instead due to the way that differently performing companies sort themselves – in or out – when deciding whether to apply for support?

Impact evaluation is essentially about *interpreting differences in a causal sense*. The challenge facing the evaluator is to avoid giving a causal interpretation to differences that are due to other factors, not to the intervention. It is necessary to identify the possible sources of bias arising in each specific situation and indicate which methods can overcome these biases, under which assumptions. *This is the essence of counterfactual impact evaluation*.

Identifying effects from before-after comparisons of beneficiaries

Let us take the first of the basic comparisons, the before-after difference. When *the same units* are observed both before and after they are exposed to an intervention, the fundamental evaluation problem is that the observed change could be due to the intervention as well as to "other changes" occurring during the same period.

The problem can be formally illustrated by the following decomposition:

$$\Delta_{B\text{-}A} = E + O_{B\text{-}A}$$

where

$\Delta_{B\text{-}A}$ Difference observed before and after the intervention, among beneficiaries

E True (and unknown) effect of the intervention on the outcome

$O_{B\text{-}A}$ Other changes occurring over time

In particular, maturation and natural evolution imply that the social or economic phenomena the intervention is trying to affect, do evolve naturally over time, in ways that are independent of the intervention. For example, the socio-economic situation of urban neighbourhoods tends to evolve over time, for better or for worse. Thus, the differences observed before and after an urban renewal programme will incorporate the (possible) effect of the programme and the results of such maturation/natural evolution.

The identification of causal effects from before-after comparisons is generally very problematic. Other than assuming away the problem—assuming temporal stability, that is, that there is no maturation or natural evolution—there is often little that can be done.

Before-after differences do not reveal the true effect of the intervention, unless we assume complete stability of other factors. Formally

$$E \textit{ is not identified by } \Delta_{B\text{-}A} \textit{ unless } O_{B\text{-}A} = 0$$

It should be stressed the different meaning of the terms "observe", "assume" and "infer". We *observe* $\Delta_{B\text{-}A}$, we can *assume* $O_{B\text{-}A} = 0$, which would allow us to *infer* that $E = \Delta_{B\text{-}A}$. The assumption $O_{B\text{-}A} = 0$ would be called the "identifying assumption", because it would be crucial in giving a casual interpretation to the observed difference.

Identifying effects by comparing beneficiaries and non-beneficiaries

By far the most common strategy to estimate the causal effect of an intervention is to exploit the fact that some "units" have been exposed to the intervention and some other have not, according to some *selection* mechanism or rule.

For example, eligible enterprises might or might not apply for state aid to finance R&D projects; unemployed workers might or might not participate in a retraining programme after a plan closure; urban neighbourhoods might or might not receive funding for urban renewal projects. Although the existence of universal policies cannot be ignored, they are relatively rare in the case on cohesion policies. In most cases, it is possible to find units that are not exposed to the policy. For simplicity, we consider only the case of a simple binary treatment, where either the units receive the treatment implied by the policy, or they do not.

The outcomes observed among beneficiaries can be compared to those among non-beneficiaries, (assuming the outcomes can be measured for both groups with the same instrument): however, this difference does not by itself reveal the true effect of the intervention on the outcome. It cannot necessarily be *interpreted* in a causal sense. The causal interpretation depends on the nature of the process that leads some units to be exposed to the intervention, while others are not.

The *observed* difference can always be thought as the sum of two components: the true effect of the policy and the difference created by the selection process itself. Neither one can actually be observed, we can only make guesses about them. The following decomposition is fundamental to show the logic behind the impact evaluation methods illustrated in this section of the Sourcebook.

$$\Delta_{T\text{-}NT} = E + S_{T\text{-}NT}$$

where

$\Delta_{T\text{-}NT}$ Difference in the outcome observed between beneficiaries and non-beneficiaries

E True (and unknown) effect of the intervention on the outcome

$S_{T\text{-}NT}$ Selection-generated differences between beneficiaries and non-beneficiaries. Keeping with the existing terminology, we will refer to the $S_{T\text{-}NT}$ as ***selection bias***.

For example, in the case of the support given to firms to invest in new equipments, the differences between the performance of supported and non-supported firms can be decomposed into the true causal effect (possibly zero) of the support and the differences due to the selection process that sorts companies into applicants and non-applicants, and then sorts applicants into recipients and non-recipients. It is very likely that supported and non-supported firms would differ in terms of performance even if the former had not received the support.

The difference observed between beneficiaries and non-beneficiaries does not reveal (identify) the true effect of the intervention unless the selection bias is zero. Formally:

$$E \textit{ is not identified by } \Delta_{T\text{-}NT} \textit{ unless } S_{T\text{-}NT} = 0$$

Again, the line of reasoning is the following: we only *observe* $\Delta_{T\text{-}NT}$, we can *assume* $S_{T\text{-}NT}= 0$, which would allow us to *infer* that $E = \Delta_{T\text{-}NT}$. Then $S_{T\text{-}NT}= 0$ would be called the “identifying assumption”.

But how does one eliminate selection bias? Eliminating selection bias represents the major challenge in conducting impact evaluations and it has received a lot of attention by the statistical, economic and sociological methodologists. A range of methods and techniques are available to (attempt to) deal with it. Knowledge of the selection process is crucial in order to choose the best methods. The methods presented in this section of the Sourcebook have a common goal: to recover the true effect of an intervention on the beneficiaries by forcing $S_{T\text{-}NT}$ to be as close as possible to zero.

The ideal strategy to eliminate selection bias: randomization

The ideal strategy to eliminate selection bias is to randomly select who becomes a beneficiary and who becomes a non beneficiary. In this case we *know* selection bias is zero. Formally:

$$E \textit{ is identified by } \Delta_{T\text{-}NT} \text{ because } S_{T\text{-}NT} = 0 \text{ by construction}$$

Unfortunately, randomization is rarely a feasible option for cohesion policies, because it requires that the control over “who received what” is given to chance. However, cohesion policies are first and foremost interventions that assign resources to local actors. Randomly assigning resources to local actors for purpose of evaluation is *politically unfeasible*, because it contradicts the very nature of the allocation process to disadvantaged areas.

At a more disaggregated level, when local actors allocate the resources to specific initiatives or projects, randomization can be used in order to learn “what works”. The learning generated by the use of randomization could motivate some local actors to adopt it despite its difficulties. Randomization as an evaluation strategy is now widely used in the context of developing countries. On the other hand, even when politically feasible, randomization still encounters many limitations (and more detractors one would expect on the basis of these limitations alone).

Randomization produces impact estimates that are internally valid, but are *difficult to generalize*: such generalization is key to the usefulness of the result for policy-making. Experiments are often costly and require close monitoring to ensure that they are effectively administered. The potential for denying treatment can pose ethical questions that are politically sensitive. These may reduce the chances of an experiment being considered as a means of evaluating a programme and may also increase the chances of those responsible for delivery of the programme being reluctant to cooperate.

Randomization requires careful planning of interventions, an early involvement of the evaluator and a degree of stability of the environment in which the experiment is taking place: all features that are rarely present in the public sector of EU Member States. Randomization requires that the intervention is fairly simple, while cohesion policies are traditionally complex, because they insist on multifaceted/multilevel problems: while complexity is an overall obstacle to evaluation, and to knowledge more generally, in the case of randomization the clash between methods and circumstances is particularly evident.

There are also practical problems that can bias the estimates. It may be that the implementation of the experiment itself alters the framework within which the programme operates. This is known as 'randomisation bias' and can arise for a number of reasons. For instance, if random exclusion from a programme demotivates those who have been randomised out, they may perform more poorly than they might otherwise have done, thus artificially boosting the apparent advantages of participation.

Another endemic problem with experiments is non compliance. This can take the form of no-shows (those assigned to treatment who drop-out before it is completed, sometimes even before it starts) or of crossovers (those assigned to control who manage to receive treatment anyway). With both no-shows and crossovers, non-experimental methods can be used to retrieve the desired parameters. However, this is a second-best position since experiments are designed specifically to avoid this sort of adjustment. Moreover, it is worth noting that the problems of programme no-shows and crossovers are not unique to experiments, although experiments may exacerbate the second problem by creating a pool of people who want to participate but were refused by the randomisation process.

To conclude, any credible strategy for evaluating the impact of cohesion policy must include in its arsenal a number of non-experimental methods and techniques (also referred to as "quasi-experimental").

The non-experimental strategies to reduce/eliminate selection bias

The general strategy pursued by the evaluator using non-experimental methods can be represented by the following expression:

$$\Delta_{T\text{-}NT} = E + ADJUST(S_{T\text{-}NT})$$

Where ADJUST is a mixture of data, institutional knowledge and arbitrary assumptions such that

$$ADJUST(S_{T\text{-}NT}) = 0 \quad \rightarrow \quad E = \Delta_{T\text{-}NT}$$

The following section illustrates four main non-experimental strategies to correct the presence of selection bias and recover the causal effect of the intervention. We examine them in turns.

Difference-in-differences in detail

The difference-in-differences identification strategy

Difference-in-differences or double differencing is based on the *precondition* that outcome data (for example, firm sales) are available for beneficiaries and non-beneficiaries (assisted and non assisted firms), both before and after the intervention (say, the year preceding and the year following the receipt of assistance).

$$\Delta_{T\text{-}NT} = E + S_{T\text{-}NT}$$

As a consequence, we also are able to observe $\Delta_{T\text{-}NT}|t\text{-}1$

Effects are obtained by subtracting $\Delta_{T\text{-}NT}|t\text{-}1$ the pre-intervention difference in outcomes between beneficiaries and non-beneficiaries from the post-intervention difference. The *identifying assumption* is that selection bias is constant in time, so that $S_{T\text{-}NT} = S_{T\text{-}NT}|t\text{-}1$.

E is identified by $\Delta_{T\text{-}NT} - \Delta_{T\text{-}NT}|t\text{-}1$ because we assume that $S_{T\text{-}NT} = S_{T\text{-}NT}|t\text{-}1$.

The result of the double difference can be interpreted as a causal effect only if the pre-post trend for non-beneficiaries is a good approximation for the (counterfactual) trend among beneficiaries. The plausibility of this assumption can be tested if more periods of pre-intervention data are available.

Description and purposes of the tool

The impact of a policy on an outcome can be estimated by computing a double difference, one over time (before-after) and one across subjects (between beneficiaries and non beneficiaries). In its simplest form, this method requires only aggregate data on the outcome variable: no covariates or microdata are strictly necessary. If sample average data is available for beneficiaries and non beneficiaries for at least two time periods, the difference-in-differences (DID) method produces estimates of impacts that are in principle more plausible than those based on a single difference (either over time or between groups). However, some untestable assumptions are still needed in order to identify impacts through double differencing.

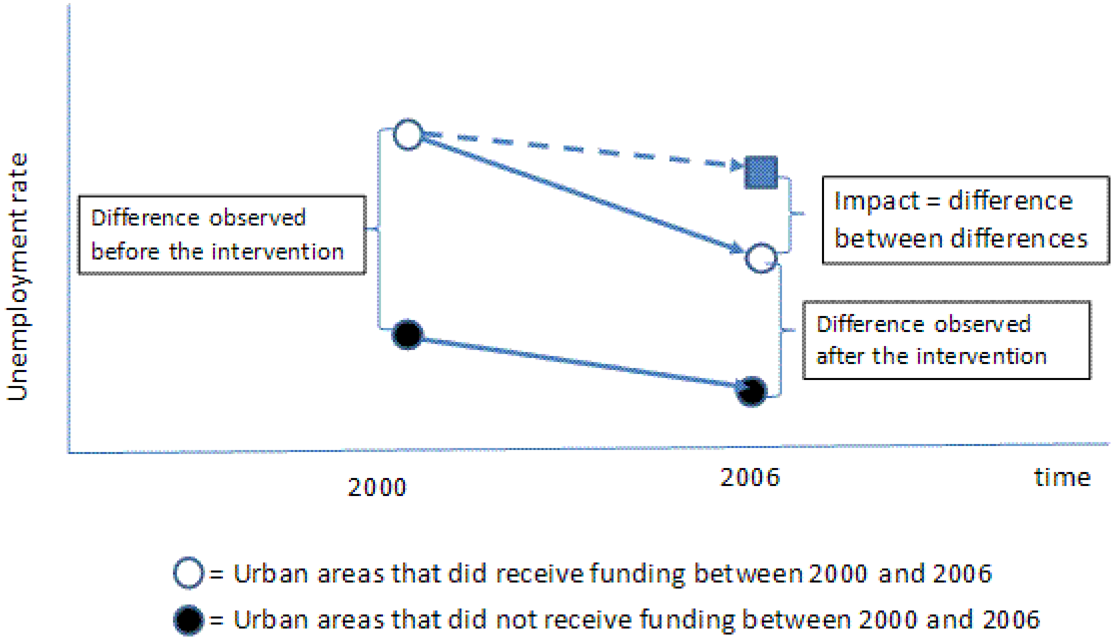
There are two ways to explain how double differencing produces impact estimates. The most intuitive is to start out with the difference in outcomes between beneficiaries and non beneficiaries, measured *after* the intervention has taken place (for example, the difference in average employment between supported and non supported SME, a year after the support has been provided.) Such difference does not reveal the effect of the intervention, since beneficiaries tend to be different from non beneficiaries even in the absence of the intervention. This is what we call *selection bias*. Now, let us suppose we have data on the outcome variable for beneficiaries and non-beneficiaries observed *before* the intervention takes place. *Subtracting* the pre-intervention difference in outcomes from the post-intervention difference eliminates one kind of selection bias, namely the kind related to time-invariant individual characteristics. In other words, if what differentiates beneficiaries and non beneficiaries is fixed in time, subtracting the pre-intervention differences eliminates selection bias and produces a plausible estimate of the impact of the intervention.

A stylized example

URBAN I and II were Community Initiatives funded through the Structural Funds, to promote regeneration in urban areas suffering from high unemployment, high levels of poverty and social exclusion, and poor environmental conditions.[1] Evaluating the success of these programmes involves answering causal questions, such as "did the urban regeneration programmes produce a positive effect on the socio-economic conditions of the areas involved?" The difference-in-differences method can provide an answer as long as the outcome of interest can be measured both before and after the implementation of the urban regeneration programme in a representative sample of both participating and non participating urban areas.

Let us take the impact on the unemployment rate: it is estimated by subtracting the difference observed between the two groups before the intervention from the difference observed after the intervention. The following picture provides a graphical illustration of this interpretation of the difference-in-difference method. On the horizontal axis we have time, with two points, one before and one after the urban regeneration initiative was implemented. Let us say, 2000 and 2006, as in the URBAN II initiative. On the vertical axis we put the unemployment rate. Each of the four circles in the graph represents an average: two are taken in 2000 and two in 2006, respectively among the 70 urban areas that received funding for urban regeneration, and among a sample of 70 comparable areas, located in the same cities, but not given any funding.[2]

Fig. 1 DID as a difference between a post-programme and a pre-programme difference


Unemployment rate
Difference observed before the intervention
Impact = difference between differences
Difference observed after the intervention
2000
2006
time
○ = Urban areas that did receive funding between 2000 and 2006
● = Urban areas that did not receive funding between 2000 and 2006

Obviously, the difference observed between the two groups of areas in 2006 is not the impact of the programme: this difference could be caused entirely by the selection process—that is, areas with higher unemployment rate had better chances of being admitted into the programme. If taken as an indication of impact, the difference shown in the graph would represent a disappointing result: that URBAN produces no useful impact on the labour market, because after the intervention the unemployment rate is higher in the funded areas than in the unfunded ones.

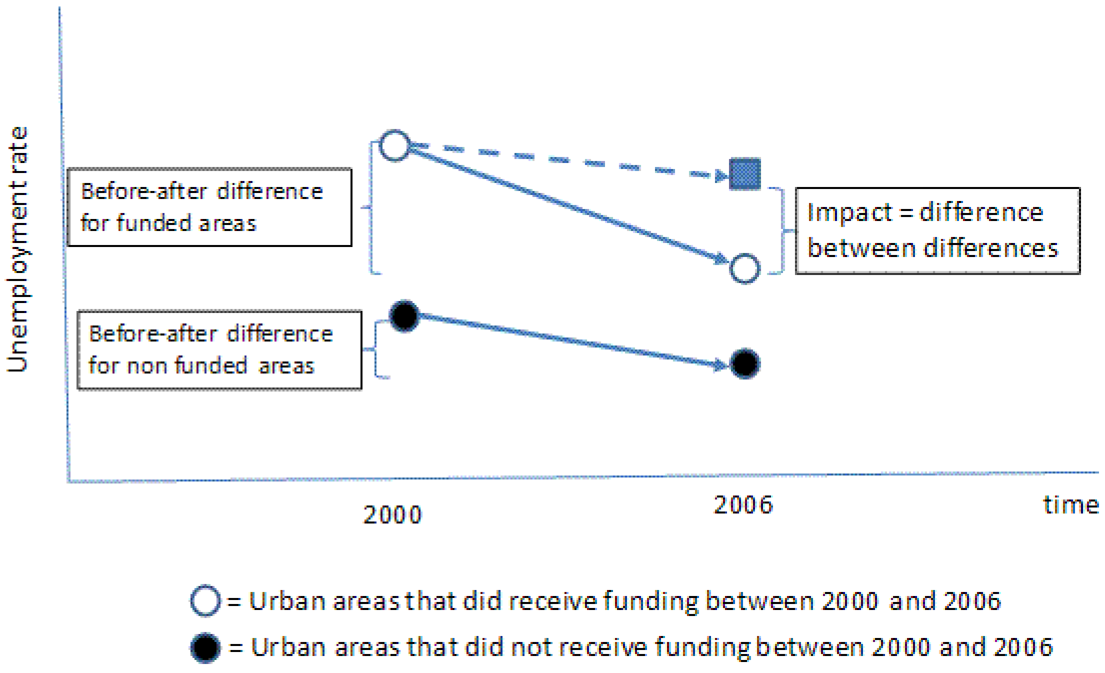
The fallacy of this interpretation is fully evident when uses data on the unemployment rate observed *before* the intervention. Figure 1 shows that in 2000 the difference in the unemployment rate between the two groups of areas was even *larger* than in 2006. It is the *reduction* in the unemployment rate gap that can be interpreted as *the impact* of the programme.

However, the validity of this conclusion depends on a crucial assumption: that in the absence of URBAN, the *trend* among funded areas would have been similar to that of the unfunded areas. Graphically, this is tantamount to drawing a dotted line parallel to the trend observed among unfunded areas, but starting where the funded areas are in 2000. This dotted line points a square in 2006: this is the counterfactual, our estimate of what would have happened to the unemployment rate in URBAN areas had URBAN *not* been implemented.

An alternative explanation

An alternative way to explain how the double differencing identifies the impact of a policy is to start from the *change observed over time among beneficiaries*. This difference *cannot* be interpreted as the impact of the policy, because many other factors and processes unfolding over time, besides the intervention, might have caused the observed change. One way to take this "natural dynamics" into account is to compute the change over time observed among non-beneficiaries during the same period. Subtracting the change observed over time among non-beneficiaries from that observed among beneficiaries produces an estimate of the impact of the programme. It is the same estimate as that shown in Figure 1, because it depends on the same crucial assumption—that in the absence of the intervention the *trend* among the two groups of areas would have been the same. This different view of the same result is illustrated in Figure 2.

Fig. 2 DID as a difference between two before-after differences


Unemployment rate
Before-after difference for funded areas
Before-after difference for non funded areas
Impact = difference between differences
2000
2006
time
○ = Urban areas that did receive funding between 2000 and 2006
● = Urban areas that did not receive funding between 2000 and 2006

The results *cannot* be different than before: the four points did not move, the dotted line is parallel to the same solid line and thus leads to the same counterfactual. What is different is the line of reasoning used to interpret the data. In the first case, one stresses selection bias and the attempt to correct it by subtracting pre-intervention differences. In the second case, one stresses the other type of distortion, due to natural dynamics, and attempts to correct it by subtracting the change observed among non-beneficiaries. In both cases, one really makes the same assumption: that of "parallelism" between what actually happened and what would have happened without the policy.

[1]The first round of the URBAN programme was launched in 1994 and ran until 1999. URBAN I supported 118 European cities in 15 Member States and had a community contribution of €950 million. Its successor, URBAN II supported 70 programmes across 14 countries and received €754 million from the European Regional Development Fund (ERDF).

[2]ECOTEC (2009), in an attempt to apply DID to the URBAN II programme, compared the unemployment rate of the URBAN II area with the rate for the city as a whole.

Circumstances in which it is applied

The applicability of the DID method requires that the outcome is *replicable* over time, that is, equivalent measurements can be taken repeatedly in successive time periods and that this repeated measurement can be done independently of the existence of the policy. Many if not most outcomes relevant for public policy are replicable over time for the same units —such as sales or profits of firms, the income of individuals or the consumption of households. We have *panel data* if the measures are taken on the same units over time.

Some outcomes have only one meaningful realization for each individual unit, such as the duration of unemployment after a job loss, or the weight of babies at birth. In these cases reliability can be obtained at a more aggregate level by using successive cohorts of individuals experiencing the same event. For example, successive cohorts of individuals entering unemployment will produce distinct estimates of the average duration of unemployment.

Another issue relevant for the applicability of DID is whether data on the outcome variable are routinely collected as part of official statistics, such as the unemployment rate and the per capita GDP, or instead outcome data must be collected ad hoc. In the latter case, a serious obstacle to the applicability of DID often comes from the fact that nobody *before* the intervention has given any thought to collecting such data, particularly at the level of geographical detail that becomes relevant after the policy is implemented.[1] If comparable pre-intervention data are lacking, one can resort to retrospective measurement, taken after the policy is implemented but with reference to both the pre-intervention period as well as the post-intervention period. The danger of such strategy is contamination between measures referring to different time periods but collected with the same interview.

The applicability of the method requires also that the intervention is of a discrete (binary) nature: one needs units that are exposed and units that are not exposed to the policy. Interventions of a continuous nature cannot be easily analysed with this method.[2]

[1]ECOTEC (2009) documents the difficulties in obtaining unemployment rate data for urban areas for the years 2000 and 2006.

[2]The reader is referred the discussion of Chapter 5 of Angrist and Pischke (2008) on many issues relevant to DID, such as a comparison with fixed-effect models, the use of covariates, as well as extensions to multiple periods and continuous treatments.

The main steps involved

In order to illustrate the steps involved a real application of the DID method will be used as an example: it is taken from an evaluation of the impact of Structural Funds in Sweden during the period 1995 to 1999 (ITPS 2004). The study was sponsored by the Swedish Institute for Growth Policy Studies and conducted by Oxford Research and the University of Umea: it consists of a "*comparison between the group of municipalities that have been recipients of structural fund projects with the group of municipalities that have not received structural funds*".

Step 1. Defining the outcome variable(s)

The analysis can be conducted with respect to as many outcome variables there are data for. The Swedish study focuses "*on the trends in three goal indicators (per capita income, employment and population) in order to see the effects the structural funds have had in the relatively poorest Swedish municipalities*". The analysis is then extended to cover intermediate outcomes, to explore the mechanisms behind the effects (or the lack thereof). We will report results for one outcome variable, the annual growth in per capita income, because the study conducts most of the analysis with respect to this variable.

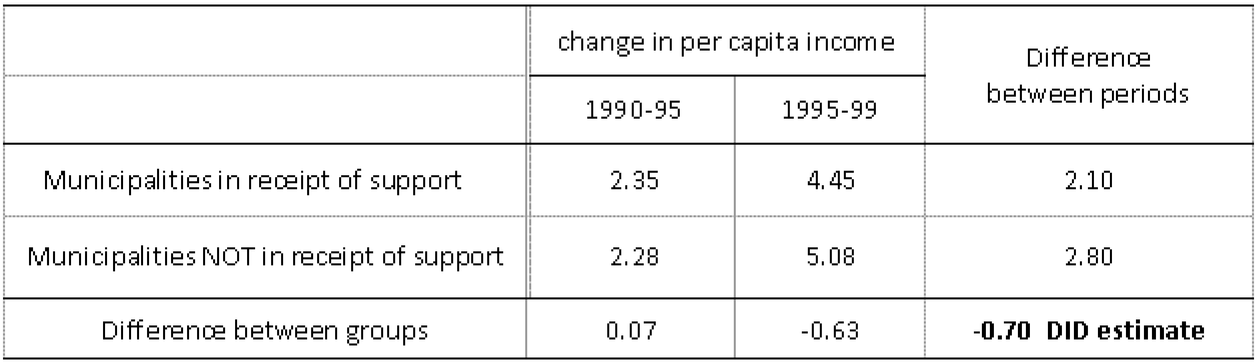
Step 2. Defining the time dimension

In the Swedish study "*the two periods that are compared are the period 1990–1995 and the period 1995–1999. The first period ends in the year the geographical programme was introduced and the second period includes the entire period of time covered by the geographical programme. The periods have been selected in such a way that they cover approximately the same length of time.*" While the latter is not a requirement, it is important that the choice of periods clearly distinguishes a "before" the intervention period and an "after" period.

Step 3. Computing the double difference

The basic analysis is simply a matter of computing averages for the two groups in the two time periods, thus obtaining a value corresponding to the four circles displayed in Figure 1 and 2. These averages are best displayed in the following format, showing the groups been compared on the rows and the time periods on the columns. The simple differences are found in the two margins, while the "difference between the differences" is shown in the lowest right cell of the table.

Tab. 1 DID estimate of the effect of support on the change in per capita income



	change in per capita income		Difference between periods
	1990-95	1995-99	
Municipalities in receipt of support	2.35	4.45	2.10
Municipalities NOT in receipt of support	2.28	5.08	2.80
Difference between groups	0.07	-0.63	**-0.70 DID estimate**

Source: The EC Regional Structural Funds impact in Sweden 1995-1999: A quantitative analysis

The table can be read in two different ways, in line with the two interpretations discussed earlier. If one reads the columns first, the focus is on the differences between the two groups of municipalities. It turns out that the two groups did not differ much in terms of per capita income growth in the five years leading up to the 95-99 structural funds intervention. They differ more sharply after the policy is enacted, in the sense that the non supported municipalities experience *higher* growth in per capita income. The DID estimate is thus the difference between an almost zero pre-intervention difference and a negative post-intervention difference, leading to a negative DID estimate.

The importance of double differencing can be more fully appreciated if one reads the table by the rows. The first row taken by itself would have one conclude that the intervention is extremely effective: an average growth rate of 2.35 percent has become a more substantial 4.45 percent—almost double. However, the other municipalities fared even better, with a 2.80 point increase in the rate of growth. The DID estimate is obviously the same as before, negative 0.70 points.

The following is the comment in the report: *"Where per capita income is concerned, the result is that the municipalities in receipt of support were the more successful of the two groups during the period 1990 to 1995 when they did not receive any support. However, during the period 1995 to 1999 municipalities in receipt of support were significantly less successful compared to municipalities not in receipt of support. Even if development trends are positive in both groups, it is the group not in receipt of support that is most successful. The difference-in-difference rating is –0.70 which shows that annual growth in municipalities in receipt of support is 0.70 percentage points lower than in municipalities not in receipt of support. This is thus a sign of an increasing difference between the two types of municipalities."*

The report is careful in not attaching a strong causal interpretation to this conclusion, talking only about "*increasing difference between the two types of municipalities*". In other parts of the report we find stronger statements, for example "*The main conclusion of the evaluation is that it is not possible to trace any effects of the EC's geographical programmes on overall regional* *development. During the period the programmes were studied, the regional differences have tended to intensify rather than be leveled out*."

It must be stressed that any causal interpretation rests on one — untestable — assumption: that in the absence of the programme the supported municipalities would have continued to enjoy the *same* growth as the non supported ones. In this particular case, this assumption seems implausible. Most likely the supported municipalities were on a lower growth path than the supported ones. If this were the case, what seems to be a *negative* impact could well turn into a zero impact, or a positive one.

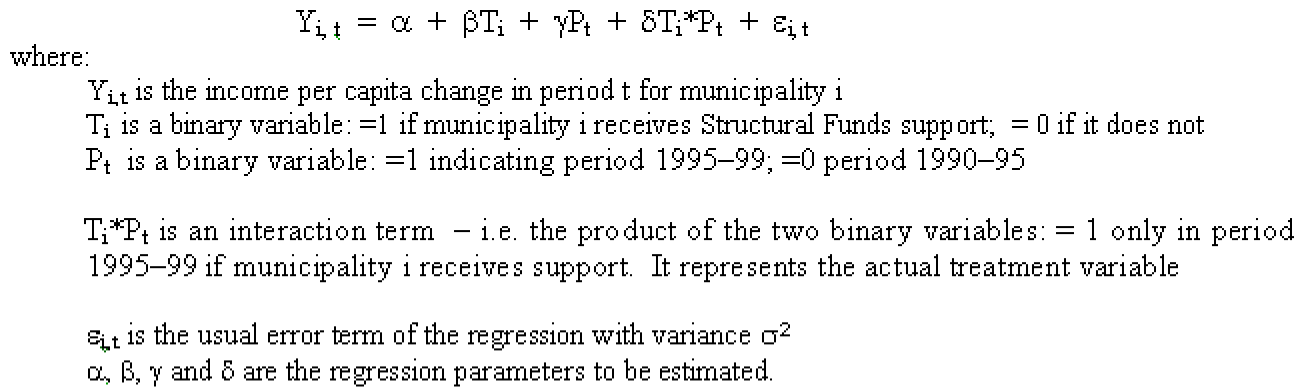
Step 4. Relaxing the assumption of "parallelism"

There are two possible extension of the simple DID method: they both require the availability of "more data" in order to relax the parallelism assumption. If one had outcome data for more time pre-intervention time periods—in the example, for the previous five years, from 1985 to 1989—one could test directly the hypothesis that the growth paths were the same in the two groups in the absence of the intervention. If it would turn out that indeed to growth paths were different, this information can be incorporated into the analysis.

The alternative to more outcome data is data on *other variables* that influence both the outcome variable and are correlated with treatment status. However, incorporating other variables entails a big loss in terms of simplicity: it requires a shift from the simple—and intuitive—differences between means to the use of a regression model estimated on microdata.

Step 5. Using regression to replicate the DID results

Let us see first how the results shown in Table 1 can be obtained through a regression model, then we will add covariates to the model. Using the same data that produced the DID estimate, one can easily estimate the following regression equation:



$$Y_{i,t} = \alpha + \beta T_i + \gamma P_t + \delta T_i{*}P_t + \varepsilon_{i,t}$$

where:

$Y_{i,t}$ is the income per capita change in period t for municipality i

T_i is a binary variable: =1 if municipality i receives Structural Funds support; = 0 if it does not

P_t is a binary variable: =1 indicating period 1995–99; =0 period 1990–95

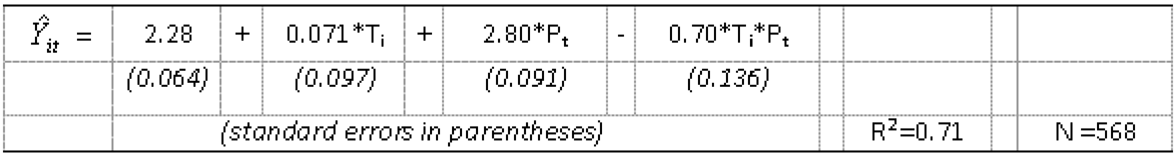
$T_i{*}P_t$ is an interaction term – i.e. the product of the two binary variables: = 1 only in period 1995–99 if municipality i receives support. It represents the actual treatment variable

$\varepsilon_{i,t}$ is the usual error term of the regression with variance σ^2

α, β, γ and δ are the regression parameters to be estimated.

The following are the estimates reported in the study:

Tab. 2 Regression estimates of the effect of support on per capita income



$\hat{Y}_{it}$ =	2.28	+	0.071*T_i	+	2.80*P_t	-	0.70*T_i*P_t				
	(0.064)		(0.097)		(0.091)		(0.136)				
	(standard errors in parentheses)								R^2=0.71		N =568

Source: "The EC Regional Structural Funds impact in Sweden 1995-1999: A quantitative analysis"

By comparing the estimates in Table 2 with those in Table 1, one can easily see that the regression exactly reproduces the estimates produced by the differences in means. More precisely, the estimate of α of 2.28 corresponds to the average income growth for the municipalities without support in the 1990-95 period. The initial difference between the two groups is reproduced by β and it is an insignificant 0.071. By contrast, very significantly different from zero is the pre-post difference for the municipalities without support, 2.80, reproduced by γ. Finally, the DID impact estimate corresponds to δ and turns out to be significant, and negative.

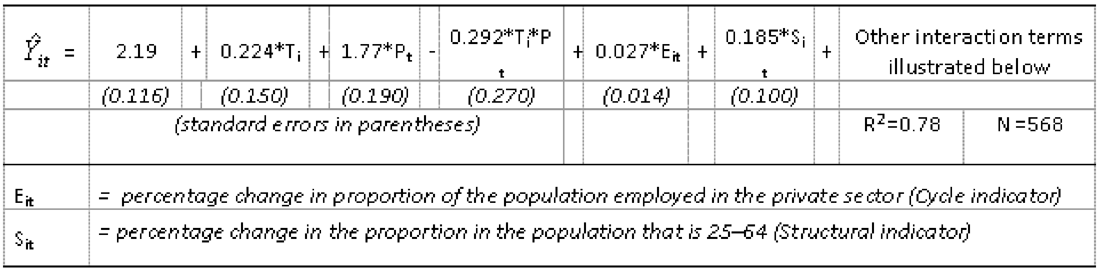
Why then go to the trouble of estimating a regression, if the results are identical to those obtained by simple differences? The main reason is that other variables can be added to the right-hand side of the equation, allowing a different way of relaxing the stringent parallelism assumption.

Step 6. Including covariates into the regression

The Swedish study adds two covariates to the regression model. One is defined as a *cycle* indicator, and it is percentage change in proportion of the population employed in the private sector in the municipality, the second is defined as a *structural* indicator and it is the percentage change in the proportion of the population aged 25–64 in the municipality. These variables are intended, according to the report, to "*test whether any periodical and/or structural changes have taken place between the two periods of time that can possibly better explain regional development than support from the EC's geographical programmes*".

The addition of the two variables to the model (including the interaction terms with the existing regressors) changes the estimates of the Structural Funds impact from negative and significant to basically zero, as shown in Table 3.

Tab. 3 Regression estimates of the effect of support on per capita income, controlling for cyclical and structural changes



$\hat{Y}_{it}$ =	2.19	+	0.224*T_i	+	1.77*P_t	-	0.292*T_i*P_t	+	0.027*E_{it}	+	0.185*S_{it}	+	Other interaction terms illustrated below	
	(0.116)		(0.150)		(0.190)		(0.270)		(0.014)		(0.100)			
	(standard errors in parentheses)												R^2=0.78	N =568
E_{it}	= percentage change in proportion of the population employed in the private sector (Cycle indicator)													
S_{it}	= percentage change in the proportion in the population that is 25–64 (Structural indicator)													

Source: The EC Regional Structural Funds impact in Sweden 1995-1999: A quantitative analysis

In the words of the Swedish report "*there are no significant differences in the extended model between the two groups of municipalities in the first period. The difference-in-difference estimate is still negative but it is not significant."*

The other interaction terms, not shown in Table 3, allow the evaluators to assess how the effects of the two explanatory variables interact with the funding in time. The report continues *"If we look more closely at the two explanatory variables, it can be seen for example that the variable proportion of private sector employees is the driving force for income growth in the group of municipalities not in receipt of support, particularly in the period 1995 to 1999, the higher economic activity in Sweden during these years seem to have benefited these municipalities."* From the complex interaction structure between the explanatory variables and the treatment and period indicators, we calculated that the effect of a percentage point increase in private employment has the following pattern:

Effect of cycle		Unfunded 1990-95	Unfunded 1995-99	Funded 1990-95	Funded 1995-99
indicator		0.027	0.308	-0.019	-0.164

It can be seen that the effect of the cycle indicator is sizeable and positive only for the municipalities not in receipt of support, in the period 1995-99.

As far as the structural indicator is concerned, the report states that *"where the variable proportion of the population in the age group 25–64 years is concerned, the picture is more diffuse. Where the municipalities in receipt of support are concerned, in the period 1990 to 1995 there was a weakly significant negative relationship between this proportion of the population and income growth, while for the period 1995–1999 there was a weakly significant positive relationship*.

Effect of		Unfunded 1990-95	Unfunded 1995-99	Funded 1990-95	Funded 1995-99
structural indicator		0.185	-0.176	-0.262	0.387

The report concludes with the following heroic explanation: "*One possible interpretation of this can be that the support disbursed during the period 1995 to 1999 has made it possible to convert a larger proportion of population in working age to growth into per capita income while this was not possible during the period during which the municipalities did not receive support*."

Strengths and limitation of the approach

Despite its wide applicability, the difference–in-differences method is not the magic bullet of impact evaluation some claim it to be. On its positive side is the fact of not requiring complex data structures to be estimated, just aggregate data on policy outcomes, collected before and after the intervention. As one applies the method in practice, its limitations start to become clear.

On the practical side, the need of pre-intervention outcome data often represents an insurmountable obstacle, most often because of lack of planning in data collection. On the more conceptual side, the simplicity of the method comes at a price in terms of assumptions: the crucial identifying assumption to obtain impact estimates is that the counterfactual trend is the same for treated and non treated units. This assumption can only be tested (and relaxed if violated), if more data are available.

In making explicit the trade-off between data and assumptions the DID method represents a great tool for teaching the logic of non-experimental methods. Its greatness is significantly reduced when the method is actually used to derive impact estimates.

Selected references

- Angrist J., Pischke J.S. [2008], *Mostly Harmless Econometrics,* Princeton University Press, NJ
- Bertrand M., Duflo E. and Mullainathan S. [2004], *How Much Should We Trust Differences-in-Differences Estimates?*, in «The Quarterly Journal of Economics», 2004, vol. 119, n. 1, pp. 249-275.
- Card D., Krueger A. [1994], *Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania*, in «The American Economic Review», 1994, vol. 84, n. 4, pp. 772-793.
- Card D., Krueger A. [1997], *Myth and Measurement: The New Economics of the Minimum Wage*, Princeton, NJ, Princeton University Press.
- ITPS [2004], *The EC Regional Structural Funds impact in Sweden 1995-1999:A quantitative analysis,* Swedish Institute for Growth Policy Studies, University of Umeå, Department of Geography.

Propensity score matching in detail

The matching identification strategy

The matching strategy is based on the possibility of observing all the relevant characteristics X of both beneficiaries and non-beneficiaries and to pick the non-beneficiaries that "look alike" beneficiaries along these characteristics.

$$\Delta_{T-NT(\text{matched on X})} = E + S_{T-NT(\text{matched on X})}$$

Once the matching is performed, the effect of the intervention is identified by the remaining difference in outcomes between beneficiaries and matched non-beneficiaries, under the assumptions that matching has also eliminated selection bias.

$$E \text{ is identified by } \Delta_{T-NT(\text{matched on X})} \text{ because } \textit{we assume that } S_{T-NT(\text{matched on X})}=0$$

The plausibility of the elimination of selection bias by matching cannot be tested: it becomes more credible as more and more X's related to the selection process are observable.

Description and purposes of the tool

The idea behind matching is simply to select a group of non-beneficiaries in order to make them resemble the beneficiaries in everything, but the fact of receiving the intervention. If such resemblance is satisfactory, the outcome observed for the matched group approximates the counterfactual, and the effect of the intervention is estimated as the difference between the average outcomes of the two groups. For example, to estimate the effect of subsidies to increase R&D spending, one matches subsidized firms with the subset of unsubsidized ones that resemble them on all the characteristics related to the selection process. The effect of the subsidy on R&D spending is estimated by the difference between average R&D spending among subsidized firms and (matched) unsubsidized ones. All this under the condition that the matching does produce two equivalent groups.

The method of matching has an intuitive appeal because by constructing a control group and using difference in means, it mimics random assignment. The crucial difference with respect to an experiment is that in the latter the similarity between the two groups covers *all* characteristics, *both observable and unobservable,* while even the most sophisticated matching technique must rely on observable characteristics *only*.

The fundamental assumption for the validity of matching is that, when observable characteristics are balanced between the two groups, the two groups are balanced with respect to all the characteristics relevant for the outcome. The larger the number of available pre-intervention characteristics, the better the chance that this assumption holds true. The existence of a substantial overlap between the characteristics of beneficiaries and non-beneficiaries (common support) is another requirement for the applicability of this method.

The curse of dimensionality and the propensity score

When performing the matching, ideally one would like to find, for each beneficiary, a non-beneficiary that is identical in all respects that are relevant in the selection process. This level of similarity is difficult to achieve, some lesser level is used in practice. A technique called propensity score matching is now commonly used.

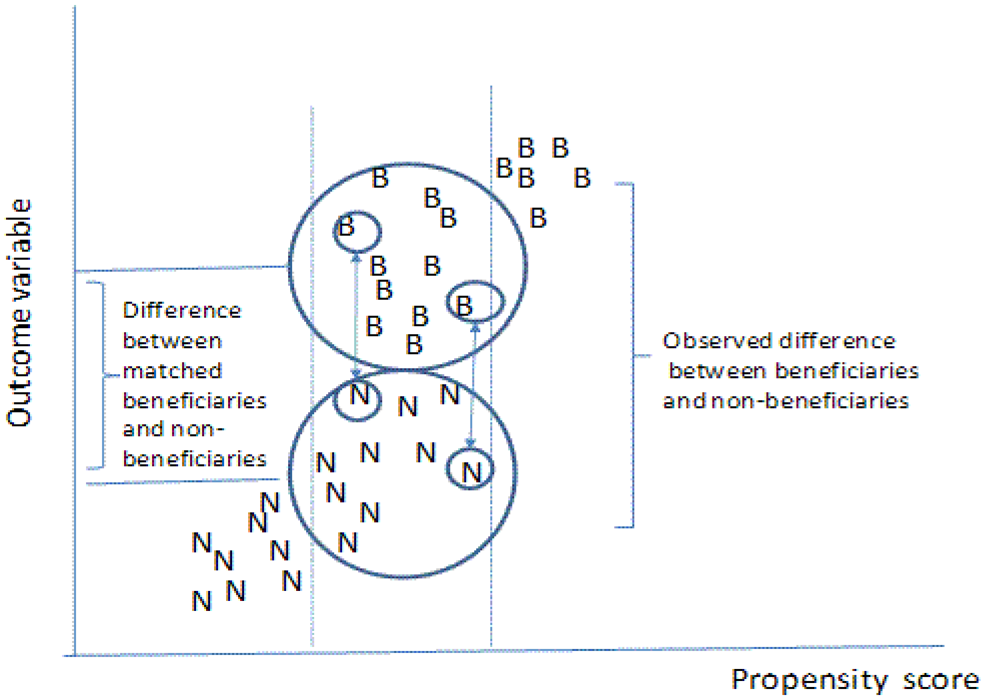
The availability of a large number of characteristics does cause a problem, known as the *curse of dimensionality*: the list of possible variables can be too large to allow a match to be achieved on each one separately, particularly if they are continuous variables. In other words, as the number of characteristics used in matching increases, the chances of finding an exact match are reduced. It is easy to see that including even a relatively small number of characteristics can quickly result in some beneficiaries remaining unmatched.

This obstacle was overcome by Rosenbaum and Rubin (1983), who suggested matching beneficiaries and non-beneficiaries solely on their 'propensity score' – the estimated probability of being a beneficiary given observable characteristics. This reduces the matching from a multi-dimensional problem (where the number of dimensions depends on the number of available variables) to a one-dimensional problem.

Intuitively, each beneficiary is matched to the non-beneficiary who is most similar in terms of probability of being a beneficiary, where this probability is calculated on the basis of individual characteristics. Once the two groups are formed, the average effect is estimated for each outcome by simply computing the difference in means between the two groups.

The following figure provides an intuitive graphical representation of the matching method. The "N" represents a sample of non-beneficiaries, while the "B" represents beneficiaries. The two dimensions of the graph are the outcome and the propensity score. In this stylized example, two matches are indicated with two small circles connected by an arrow. The important message of the picture is that, by matching on the propensity score, we obtain a difference between the means of matched beneficiaries and non-beneficiaries which is substantially different from the difference between the means of all beneficiaries and non-beneficiaries.

Figure 1. A graphical representation of matching on the propensity score


Outcome variable
Propensity score
Difference between matched beneficiaries and non-beneficiaries
Observed difference between beneficiaries and non-beneficiaries
B
N

This section draws heavily on Caliendo and Kopeinig (2008) and on Bryson, Dorsett and Purdon (2002), to which the reader is referred to for further details.

Circumstances in which it is applied

A crucial condition for the applicability of matching is the availability of characteristics observed before the intervention takes place. Variables observed after the intervention could themselves be influenced by the intervention. Ideally, all variables affecting the selection process should be included in the list of matching variables, although this is rarely the case.

Another condition for the correct application of matching is the existence of a substantial overlap between the characteristics of beneficiaries and non-beneficiaries. This overlap is known as “common support”. The most intuitive representation of the common support problem is given in Figure 1 above: the area between the two vertical bars, in which one can find beneficiaries and non-beneficiaries sharing similar values for the probability of being exposed to the intervention, is the common support. Those non-beneficiaries with very low propensity score to the left of the left-most line, and those beneficiaries with very high propensity score to the right of the right-most line, should be excluded from the comparison.

Most of the application of matching is to policies with a binary treatment, so that beneficiaries and non-beneficiaries form two separate groups. Most available methods indeed apply to this situation only. However, some recent developments extended the matching to the case in which treatment assumes several values and to that in which treatment is a continuous variable (Hirano and Imbens, 2004).

When pre-intervention outcomes are available for both beneficiaries and non-beneficiaries, matching can be combined with the difference-in-differences approach: first, units are matched on pre-intervention characteristics; the effects are then estimated by double differencing. The joint application of the two methods increases the chances that selection bias has indeed been eliminated.

The main steps involved

There are four main steps involved in the application of statistical matching to impact evaluation: estimating the propensity score, matching the units using the propensity score, assessing the quality of the match, and estimating the impact and its standard error.

To make the discussion more concrete, a real example will be illustrated in some detail, taken from an evaluation of the Regional Development Grant (RDG), a Swedish business support programme. The capital subsidies cover up to 35% of an investment. For a firm to be eligible for a subsidy, it must be used for investments in machinery, equipment, buildings or a service activity that aims to increase the market for the enterprise. Before approval of an application for support it is assessed by the county administrative board. Larger support, which exceeds 25 million SEK, is granted by the Swedish agency for economic and regional growth (NUTEK).

The objective of the study is to investigate whether firms who have received the Regional Development Grant are performing better than those firms that have not received the subsidy. The information is drawn from Gadd, Hansson and Månsson (2008).

Step 1: Estimating the Propensity Score

First a decision has to be made concerning the estimation of the propensity score. One has not only to decide about the probability model to be used for estimation, but also about variables which should be included in the model. In principle any discrete choice model can be used. Preference for logit or probit models over the linear probability model derives from the shortcomings of the latter, that can produce predicted probabilities outside the [0; 1] bounds. Logit and probit models usually yield similar results: hence, the choice is not too critical.

More attention is required regarding the inclusion of variables in the propensity score model. The matching strategy builds on the assumption that, conditional on the propensity score, selection bias is eliminated. Hence, implementing matching requires choosing a set of variables that credibly satisfy this condition. Omitting important variables can seriously increase bias in the resulting estimates of the impact of the policy. Only variables that simultaneously influence the participation decision and the outcome variable should be included. Hence, a sound knowledge of previous research and also knowledge of the institutional settings should guide the researcher in building up the model. For example, the list of variables that affects both the probability of being subsidized and the outcome (we used profits) may include size (in terms of turnover and employment), legal status, sector, market share and degree of unionization.

It should also be clear that only variables that are unaffected by participation should be included in the model. To ensure this, variables should either be fixed over time or measured before participation. In the latter case, it must be guaranteed that the variable has not been influenced by the anticipation of participation.

Finally, the data for beneficiaries and non-beneficiaries should stem from the same sources (e.g. the same questionnaire). In cases of uncertainty of the proper specification, sometimes the question may arise if it is better to include too many rather than too few variables. Although the inclusion of non-significant variables will not bias the estimates, it can increase their variance.

The evaluation of the Regional Development Grant includes in the model a range of firm characteristics and regional contextual variables [1]. These variables are intended to capture the primary information the subsidy administrator considers and would be considered by commercial banks and other institutions granting loans: variables referring to the company's profitability and financial position, the rate of return on total assets (ROA), shareholders' equity divided by total assets, the number of employees in1999, and a variable that equals 1 if the company is primarily a manufacturing company and 0 otherwise. At the municipality level the model includes the characteristics important to receive subsidy, divided into four areas: composition of the residents, location based characteristics, economic variables such as unemployment rate, income, migration, share state employed, and political situation.

Step 2: Matching the units using the propensity score

Once the propensity score model is estimated and the score computed for each unit, the next step consists of performing the actual matching—that is, after choosing a matching algorithm. Matching algorithms differ not only in the way the neighbourhood for each treated individual is defined, but also with respect to the weights assigned to these neighbours. We will not discuss the technical details of each estimator here but rather present the general ideas and the trade-offs involved with each algorithm.

a. Nearest Neighbour Matching

The most straightforward matching estimator is nearest neighbour matching. One individual from the comparison group is chosen as a matching partner for a treated individual that is closest in terms of propensity score.

Two variants are *possible*, matching `with replacement' and `without replacement'. In the former case, an untreated individual can be used more than once as a match, whereas in the latter case it is considered only once. Replacement involves a trade-off between bias and precision: if we allow replacement, the average quality of matching will increase and the bias will decrease, but fewer cases will be used, reducing precision. This issue is of particular interest with data where the propensity score distribution is very different in the treatment and the control group. For example, if we have a lot of treated individuals with high propensity scores but only few comparison individuals with high propensity scores, we get bad matches as some of the high-score participants will get matched to low-score non-participants. This can be overcome by allowing replacement, which in turn reduces the number of distinct non-participants used to construct the counterfactual outcome and thereby increases the variance of the estimator.

A problem related to matching without replacement is that estimates depend on the order in which observations get matched. Hence, when using this approach it should be ensured that ordering is randomly done.

b. Calliper and Radius Matching

Nearest neighbour matching faces the risk of bad matches, if the closest neighbour is far away. This can be avoided by imposing a tolerance level on the maximum propensity score distance (calliper). Imposing a calliper works in the same direction as allowing for replacement. Bad matches are avoided and hence the matching quality rises. However, if fewer matches can be performed, the variance of the estimates increases.

Applying calliper matching means that an individual from the comparison group is chosen as a matching partner for a treated individual because it lies within the calliper (`propensity range') and is closest in terms of propensity score. A possible drawback of calliper matching is that it is difficult to know a priori what choice for the tolerance level is reasonable.

A variant of calliper matching is called radius matching. The basic idea of this variant is to use not only the nearest neighbour within each calliper but all of the units within the calliper. A benefit of this approach is that it uses only as many comparison units as are available within the calliper and therefore allows for usage of extra (fewer) units when good matches are (not) available.

The study of the Regional Development Grant uses calliper matching "*Since the matching consists of few treated in relation to many controls, we employ the Radius matching technique which use all the comparison members within a certain boundary or radius (the calliper).Radius matching and other over-sampling techniques such as kernel matching are to be recommended when the control group is large and there is more than one nearestneighbour*."

c. Stratification Matching

Stratification matching is based on a simple idea: rank both treated and controls on the basis of their propensity score, and then group them into K intervals (strata). One stratum will contain the observations with the lowest, say, quintile of the propensity scores, the next stratum the observations with higher propensity score, and so on to the highest values. The number of treated and controls will differ from one stratum to the next, and typically the strata with lower propensities contain more controls and fewer treated, while the opposite is true for higher strata. If the number might vary, the average propensity within each stratum must not differ systematically between treated and controls, and this must be confirmed by a standard t-test (balancing).

Once the stratification is done and the balancing satisfied, we calculate the impact for each k-th stratum, simply by taking the mean difference in outcomes between treated and control observations for that stratum. The overall impact is then obtained by calculating a weighted average of the strata effects, with weights proportional to the number of treated units in each stratum. If in a stratum there are no controls, the observations in that interval will be eliminated from the analysis, in the sense that the stratum gets zero weight. One question is how many strata should be used. Five subclasses are often enough to remove 95% of the bias associated with one single covariate, as first shown by Cochrane and Chambers (1965).

d. Kernel Matching

The first two matching algorithms discussed above have in common that only some observations from the comparison group are used to construct the counterfactual outcome of a treated individual. Kernel matching uses weighted averages of *all* individuals in the control group to construct the counterfactual outcome. Weights depend on the distance between each individual from the control group and the participant observation for which the counterfactual is estimated. The kernel function assigns higher weight to observations close in terms of propensity score to a treated individual and lower weight on more distant observations. One major advantage of this approach is the lower variance which is achieved because more information is used, while a drawback is that possibly bad matches are included. Nearest neighbour can be considered an extreme form of kernel matching, where all the weight is given to the closest propensity.

How to choose: trade-offs in terms of bias and efficiency

Having presented the different possibilities, the question remains of how one should select a specific matching algorithm. In small samples the choice of the matching algorithm can be important, where usually a trade-off between bias and variance arises. So what advice can be given to the evaluator facing the problem of choosing a matching algorithm? It should be clear that there is no `winner' for all situations and that the choice of the estimator crucially depends on the situation at hand. The performance of different matching estimators varies case-by-case and depends largely on the data structure at hand. To give an example, if there are only a few control observations, it makes no sense to match without replacement. On the other hand, if there are a lot of comparable untreated individuals it might be worth using radius/calliper matching. Pragmatically, it seems sensible to try a number of approaches. Should they give similar results, the choice may be unimportant. Should results differ, further investigation may be needed in order to reveal more about the source of the disparity.

Step 3. Assessing the Quality of the Match

The next step is to check the common support between treatment and control group. The most straightforward way is a visual analysis of the density distribution of the propensity score in both groups. Lechner (2000) argues that given that the support problem can be spotted by inspecting the propensity score distribution, there is no need to implement a complicated formal estimator. However, some formal guidelines might help the researcher to determine the region of common support more precisely.

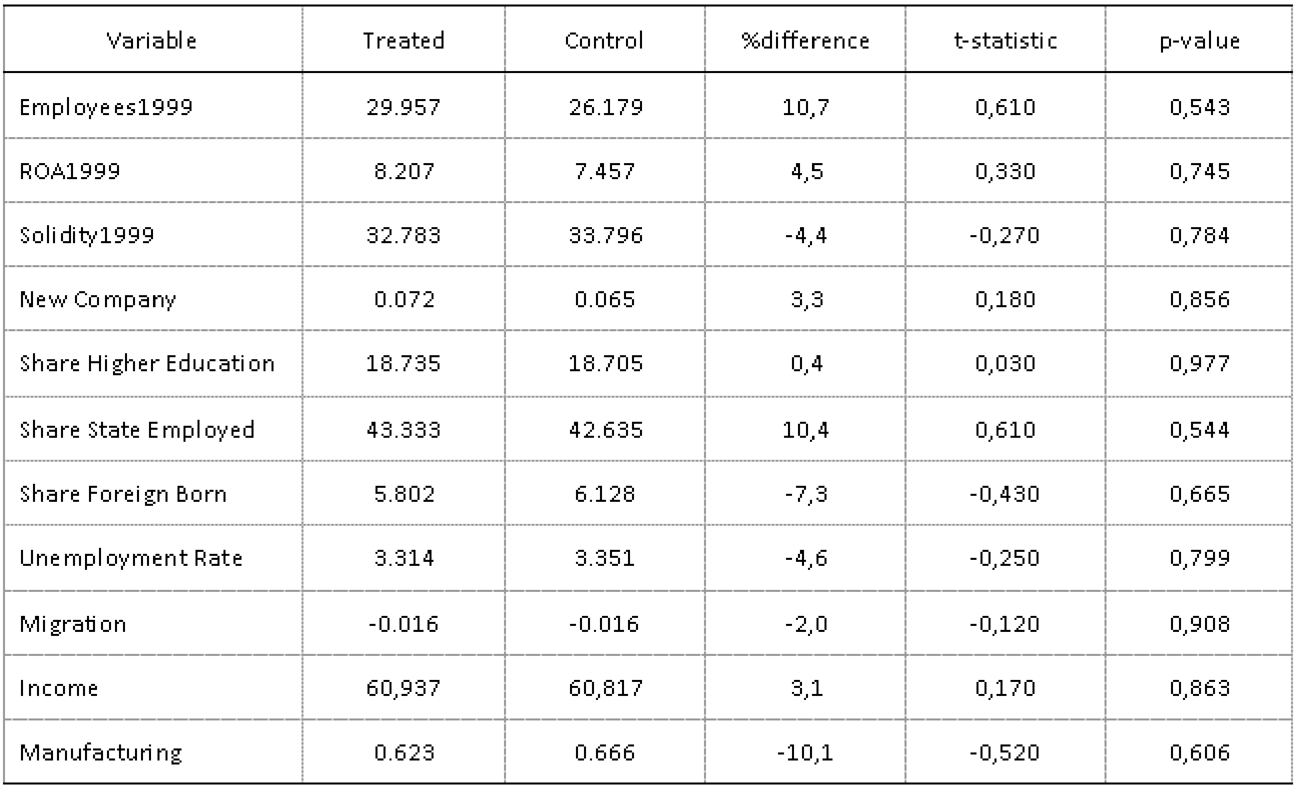
One possible method is based on comparing the minima and maxima of the propensity score in both groups. All observations whose propensity score is smaller than the minimum and larger than the maximum in the opposite group are deleted. To give an example let us assume for a moment that the propensity score lies within the interval [0:10; 0:95] in the treatment group and within [0:05; 0:90] in the control group. Hence, with the `minima and maxima criterion', the common support is given by [0:10; 0:90]. Observations which lie outside this region are discarded from the analysis.

Once one has defined the region of common support, individuals that fall outside this region have to be disregarded and for these individuals the treatment effect cannot be estimated. When the proportion of lost individuals is small, this poses few problems. However, if the number is too large, there may be concerns whether the estimated effect on the remaining individuals can be viewed as representative. It may be instructive to inspect the characteristics of discarded individuals since those can provide important clues when interpreting the estimated treatment effects.

Since we do not condition on all covariates but on the propensity score, it has to be checked if the matching procedure is able to balance the distribution of the relevant variables in both the control and treatment group. There are several procedures for this. The basic idea of all approaches is to compare the situation before and after matching and check if there remain any differences after conditioning on the propensity score. If there are differences, matching on the score was not (completely) successful and remedial measures have to be done, e.g. by including interaction-terms in the estimation of the propensity score.

A rather simple method is employed by the Swedish study on Regional Development Grant. The authors simply compute the percentage difference between mean treated and mean control (Table 1). An examination of the means of the treated and matched control group reveals that the two groups indeed seem similar. The third column lists the percentage bias between the treated and the control. The fourth column lists the results of a t-test of the equality of means between the treated and the control. With p-value well above 0.1 indicates that the null hypothesis of equal means cannot be rejected at the 10 percent level for all variables. The authors conclude "*we therefore be confident in that the results concerning the differences between the treated and the untreated firms are based on similar firms.*"

Table 1: Assessing the Quality of the Match in the Swedish study on Regional Development Grant



Variable	Treated	Control	%difference	t-statistic	p-value
Employees1999	29.957	26.179	10,7	0,610	0,543
ROA1999	8.207	7.457	4,5	0,330	0,745
Solidity1999	32.783	33.796	-4,4	-0,270	0,784
New Company	0.072	0.065	3,3	0,180	0,856
Share Higher Education	18.735	18.705	0,4	0,030	0,977
Share State Employed	43.333	42.635	10,4	0,610	0,544
Share Foreign Born	5.802	6.128	-7,3	-0,430	0,665
Unemployment Rate	3.314	3.351	-4,6	-0,250	0,799
Migration	-0.016	-0.016	-2,0	-0,120	0,908
Income	60,937	60,817	3,1	0,170	0,863
Manufacturing	0.623	0.666	-10,1	-0,520	0,606

Source: Gadd, Hansson and Månsson (2009) "Evaluating the impact of firm subsidy using a multilevel propensity score approach"

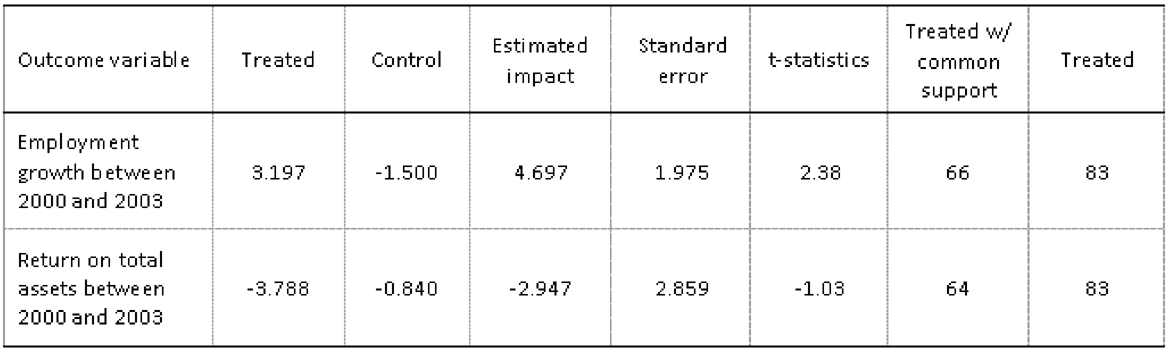
Step 4: Estimating the Average Effect and its Standard Error

After the match has been deemed of good quality, computing the effect becomes the easy task: it is enough to compute the sample averages of the two groups and calculate the difference. Before running a t-test to check the statistical significance of the effect, however, one needs to compute standard errors. This is not a straightforward thing to do.

The problem is that the estimated variance of the treatment effect should also include the variance due to the estimation of the propensity score, the imputation of the common support, and possibly also the order in which treated individuals are matched. These estimation steps add variation beyond the normal sampling variation.

One way to deal with this problem is to use bootstrapping as suggested by Lechner (2002). This method is a popular way to estimate standard errors in case analytical estimates are biased or unavailable. Even though Imbens (2004) notes that there is little formal evidence to justify bootstrapping, it is widely applied.

Table 2: The main impact estimates in the Swedish study on Regional Development Grant



Outcome variable	Treated	Control	Estimated impact	Standard error	t-statistics	Treated w/ common support	Treated
Employment growth between 2000 and 2003	3.197	-1.500	4.697	1.975	2.38	66	83
Return on total assets between 2000 and 2003	-3.788	-0.840	-2.947	2.859	-1.03	64	83

Source: Gadd, Hansson and Månsson (2009) "Evaluating the impact of firm subsidy using a multilevel propensity score approach".

The authors comment on the first result: "*The positive difference for the difference in employees indicates that the companies which received the RDG subsidy increased their number of employees more than their matched companies in the control group. This positive difference is also significantly different from zero, which can be seen from Column 6 where the t-statistics are presented, … this is to be considered a relatively large change in employment growth*."

On the other hand, the results for return on total assets are far from being significant. This suggests that the companies which received the RDG subsidy had neither a higher nor lower return on total assets, compared to the matched companies in the control group.

The authors conclude: "*The results concerning the effects of the RDG subsidy are mixed: a positive effect with respect to employment, but no effect concerning return on total assets. This result is in accordance with previous results concerning the effectiveness of the RDG subsidy. Examining the period 1990-1999, ITPS found that RDG had some effect on employment for certain periods, but did not find any effects concerning return on total assets*".

[1] Sweden was divided into 21 counties and into 289 municipalities in the year 2000.

Strengths and limitations of the approach

Matching has two clear disadvantages relative to experimental techniques. The first is the need to assume conditional independence—that is, that selection bias is eliminated by controlling for observables. In the case of properly conducted random assignment, we can be confident that the beneficiary and non-beneficiary populations are similar on both observable and unobservable characteristics. Second, whereas matching can only estimate treatment effects where there is overlap between beneficiary and non-beneficiary populations, random assignment ensures that there is common support across the whole sample. These considerations make experimental techniques unambiguously superior to matching. However, practical considerations are also important in the design and execution of programme evaluations and often these practical considerations favour matching over random assignment.

Matching's main advantage over random assignment is that it avoids the ethical considerations which arise when a potentially beneficial treatment is denied at random. Cost is also an important practical consideration when conducting evaluations. In some cases, despite matching's onerous data requirements, data generation may be less costly than in the case of an experiment since the latter involves substantial monitoring to ensure random allocation.

What are the advantages of matching relative to other non-experimental evaluation techniques? Matching is generally preferred to standard regression methods for two reasons. First, matching estimators highlight the problem of common support. Where there is poor overlap in support between the beneficiaries and non-beneficiaries, this raises questions about the robustness of traditional methods. Secondly, matching does not require functional form assumptions for the outcome equation. Regression methods impose a form on relationships (usually linear) which may or may not be accurate and which matching avoids: this is valuable since these functional form restrictions are usually justified neither by theory nor the data used (Angrist and Pischke, 2008).

In common with most other quantitative evaluation techniques, matching is not particularly well-suited to 'getting inside the black box' of how a programme does or does not work. That said, it is capable of revealing dimensions along which selection into the programme has occurred (through the participation equation and enforcing common support), and it can generate programme effects for sub-groups which can also indirectly cast light on the way the programme operates.

Selected References

Angrist J, Pischke J.S. [2008], *Mostly Harmless Econometrics,* Princeton University Press, NJ.

Bryson A., Dorsett R. and Purdon S. [2002], *The Use of Propensity Score Matching in the Evaluation of Active Labour Market Policies*, UK Department for Work and Pensions, London.

Caliendo M., Kopeinig S. [2008], *Some Practical Guidance for the Implementation of Propensity Score Matching*, «Journal of Economic Surveys», 22(1), pp. 31-72.

Cochrane W., Chambers S. [1965],*The Planning of Observational Studies of Human Populations*, in «Journal of the Royal Statistical Society», Series A, 128, pp. 234-266.

Gadd H., Hansson G. and Månsson J. [2008], *Evaluating the Impact of Firm Subsidy Using a Multilevel Propensity Score Approach*, Centre for Labour Market Policy Research (CAFO), School of Management and Economics, Växjö University.

Hirano K., Imbens G. [2004], *The Propensity Score with Continuous Treatments,* in Missing Data and Bayesian Methods in Practice: Contributions by Donald Rubin's Statistical Family, Wiley, NYC.

Imbens G. [2004], *Nonparametric Estimation of Average Treatment Effects under Exogeneity: A Review*, in «The Review of Economics and Statistics», 86(1), pp. 4-29.

Lechner M. [2002], *Some practical issues in the evaluation of heterogeneous labour market programmes by matching methods*, in «Journal of the Royal Statistical Society», Series A, 165, pp. 59-82.

Rosenbaum P., Rubin D. [1983], *The Central Role of the Propensity Score in Observational Studies for Causal Effects*, in «Biometrika», 70, pp. 41-50.

Discontinuity design in detail

The discontinuity identification strategy

The strategy is based on the idea of *discontinuity in treatment around a threshold,* which applies mainly to those situations in which some units are made eligible for the intervention and others are made ineligible by some well defined rule, typically some administrative rule.

The two groups are similar in other respects, but they are (sharply) divided according to their position with respect to a threshold, indicated with C*: those on one side of the threshold are exposed to the policy, those on the other side are not.

The essential idea for identifying the effect is that around C* with has a situation similar to randomization. Let us indicate with $\approx C^*$ a neighbourhood of C*. Forma

$$\Delta_{T\text{-}NT} \approx C^* \quad = \quad E \approx C^* \quad + \quad S_{T\text{-}NT} \approx C^*$$

The effect of the treatment (around the threshold) is obtained by the difference in outcomes around the threshold. The identifying assumption (more credible than most) is that selection bias is zero around the threshold.

$$E \approx C^* \;\; \textit{is identified by} \;\; \Delta_{T\text{-}NT} \approx C^* \;\; \text{because} \; \textit{we assume that} \; S_{T\text{-}NT} \approx C^* \; = 0$$

It should be noted that the estimated effect is a *local* effect: it is more credible (internal validity) but less generalizable (external validity).

Description and purposes of the tool

This method is applicable when the eligibility for a programme is determined by a rule of the following type: those above a certain threshold are eligible for the programme, while those below are not eligible (or *vice versa*). The threshold is a cut-off score on a continuous variable—for example age for persons, income for households and number of employees for enterprises.

Estimates of programme impact can be obtained by comparing *marginal* participants and *marginal* non-participants. The term marginal refers to the units not too far from the threshold for selection, on either side. In the neighbourhood of the threshold we have a situation that resembles randomization: the units around the threshold receive sharply different treatments, despite having similar values for the selection variable: this “unequal treatment of equals” represents the source of the identification of the impact.[1] This allows one to obtain impact estimates without imposing any other assumption, and this is the major strength of the method. However, comparing marginally treated and marginally excluded units identifies the impact of the intervention only “locally”, and not for the entire population, and this is the major weakness of the method.

The estimates obtained around the threshold cannot be generalized without making assumptions regarding the relationship between the selection variable and the outcome—that is, without some kind of regression model. This explains the odd name by which the method is known, “Regression Discontinuity Design”, a name that says more about its original understanding than its true essence. Rather than Regression Discontinuity Design (RDD), we would prefer to call this method Discontinuity of Treatment around a Threshold (DTT) but the reader would not find such a term in the published literature. The first application of RDD can be traced to Thistlethwaite and Campbell’s (1960), who estimated the effect of receipt of a National Merit Award on a student’s later success. As the award was given to students who achieved a minimum score, differences in subsequent academic achievement between those students above and below that cut-off was attributed to the effect of the award. This method is enjoying a renewed interest in programme evaluation as, according to some, the best alternative to randomization (Cook, 2008).

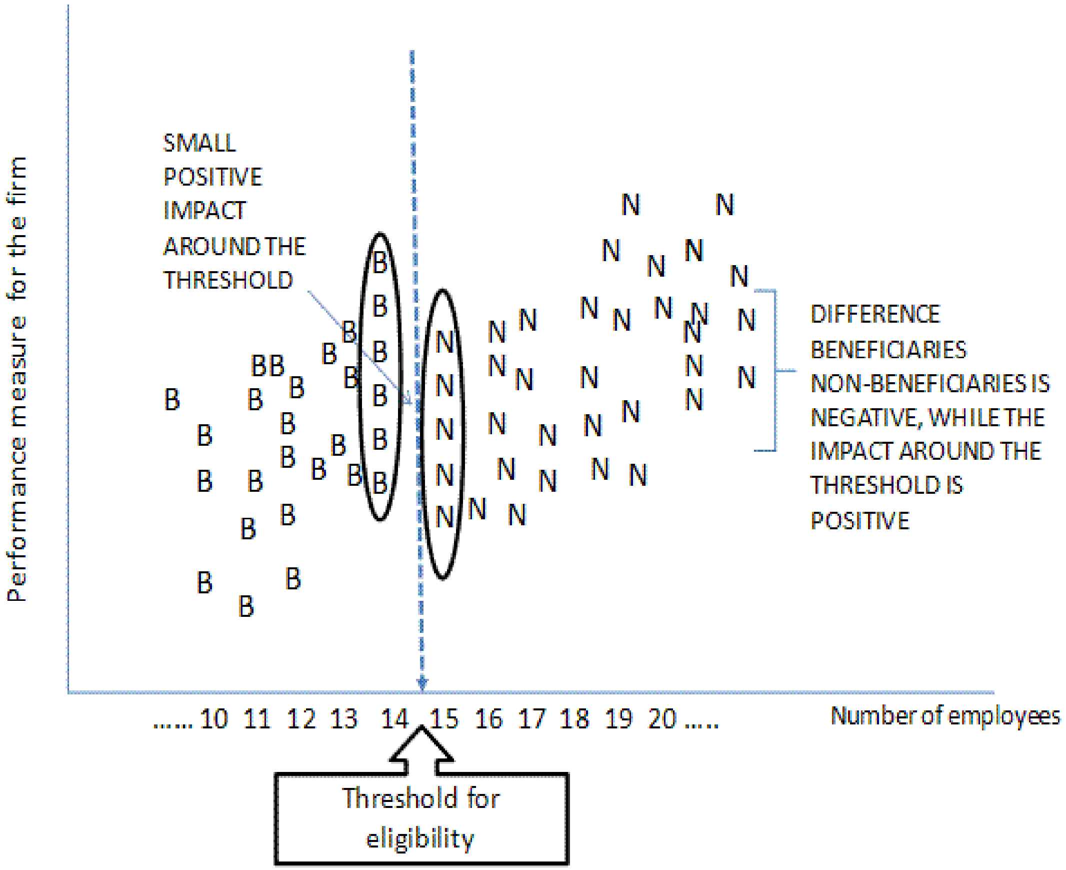
An intuitive example

Consider a policy that provides investment subsidies to small enterprises, defined as those with less than 15 full-time employees: the aim is to improve the firms' performance in some dimension, such as sales or profits. The situation in presented in Figure 1.

Firms to the left of the threshold (the B's) have less than 15 employees and are eligible for the subsidy. For the time being, we assume that all eligible take advantage of the subsidy. Firms to the right of the threshold (the N's) do not have access to the subsidy, and this rule is strictly enforced. Larger firms tend to perform better on the performance measure, and smaller firms are targeted for the subsidy. Thus the intervention has a clear *compensatory rationale;* it targets the less advantaged fraction of the population. The results of naive comparisons can be misleading: if one were to take all the N's and all the B's and compare their average performances, one would conclude that the subsidy *lowers* firms' performance. We know that such causal interpretation is unwarranted because the two groups are different even in the absence of the intervention (and there is negative selection bias).

The RDD method does not suffer from selection bias: if we restrict the attention to marginal individuals and compare firms with 14 employees (marginally eligible) to firms with 15 employees (marginally ineligible), it seems reasonable to assume that a difference of 1 employee will have a minor impact on the performance measure. By contrast, the two subgroups are treated very differently by the programme, so that around the threshold we have a randomization of sorts. The difference between the average performances of the two samples of 14-employee firms and 15-employee firms is a credible estimate of the impact of the programme. In the example, the impact around the threshold is positive, indicating that the programme works. Visually, the impact is represented by the "downward slide" of the N's just above the threshold with respect to the B's right below the threshold.

Fig.1 Comparison around a threshold (for an intervention that has a compensatory nature)


Performance measure for the firm
SMALL POSITIVE IMPACT AROUND THE THRESHOLD
DIFFERENCE BENEFICIARIES NON-BENEFICIARIES IS NEGATIVE, WHILE THE IMPACT AROUND THE THRESHOLD IS POSITIVE
B
N
...... 10 11 12 13 14 15 16 17 18 19 20 .....
Number of employees
Threshold for eligibility

By construction, these estimates are applicable only to the firms around the threshold: if impacts vary considerably with the size of the firm, the estimates of impacts cannot be easily generalized beyond firms situated around the threshold. Such generalization would require some assumption on the relationship between firm size and the performance measure—thus, some kind of regression model.

[1] As a matter of fact, some kind of "unequal treatment of equals" is at the basis of all identification strategies: equality is forced through statistical adjustment, obtained fully by randomization, obtained marginally by discontinuity, obtained partially by differencing out in time or involuntary variation.

Circumstances in which it is applied

This method has to meet several conditions for full applicability. First and foremost, selection must be determined by the position with respect to a threshold, defined along a continuous variable. Example of administrative rules of this kind are not uncommon: in addition to characteristics such as income for households or size of firms, one can think of rankings attributed by peer-review panels, score assigned by procurement committees, measures of duration or cumulated time in a given status, such as unemployment or job tenure.

Another restriction on the applicability of RDD is that the individuals should not be able to manipulate their position with respect to the threshold in order to participate in the programme. This problem is known as the "manipulation of the covariate". A problem of this sort is likely to arise in the example just given, since the position of the enterprise with respect to the threshold can be manipulated by the enterprise itself, who might be influenced by grant eligibility in their hiring decisions. The major advantage of RDD, creating a situation similar to randomization, would be lost if firms interested in the subsidy would “pile-up” at the 14 employee mark. The occurrence of this event is testable, because it translates into a spike in the frequency distribution of firm size.

Another concern about RDD designs is the possibility of other changes occurring at the same cut-off value of the covariate. Such changes may affect the outcome, and these effects may be attributed erroneously to the treatment of interest. In the example just given, this problem would appear if “15-employees” were also the cut-off point, for example, for the applicability of labour-protection legislation or for eligibility for unemployment benefits. One would not be able to disentangle which policy creates the change around 15 employees.

Sharp and fuzzy design

Finally, in its purest form, this method requires a “sharp” discontinuity in treatment around the threshold: the probability of treatment should go from zero for those on one side to 1 for those on the other side. Formally:

P(treatment|below the threshold)=0 and P(treatment|above the threshold)=1

This is often not the case. In the above example, a fraction of small firms might not take advantage of the subsidy. Thus moving across the threshold the treatment varies from zero to some number larger than zero but less than one. Such a situation can arise if incentives to participate in a programme change discontinuously at a threshold, without these incentives being powerful enough to move all units from nonparticipation to participation. Thus the discontinuity in treatment becomes “fuzzy”.

P(treatment|below the threshold) < P(treatment|above the threshold)

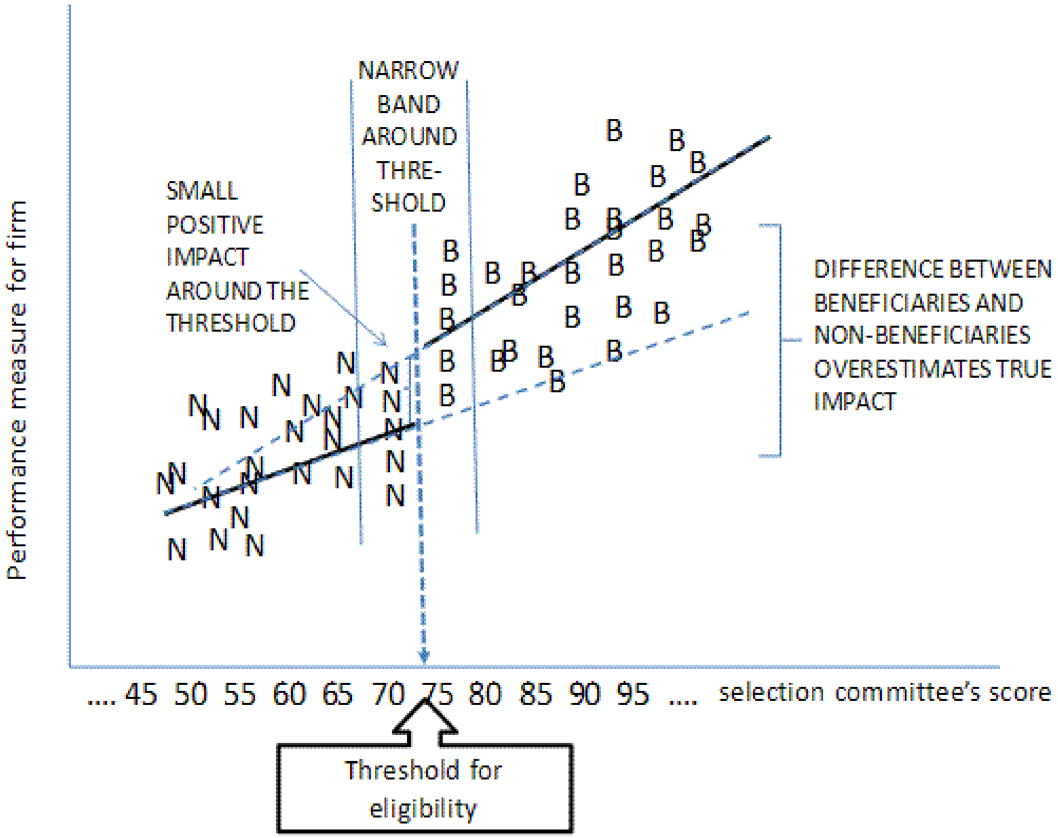
We consider first the sharp RDD, and delay discussion of fuzzy RDD to the next section.

Another intuitive example, adding the regression:

We provide a second stylized example, in which the three limitations just mentioned (manipulation of the covariate, coincidence with other cut-off points, fuzzy discontinuity) do not apply. It is the case of an investment subsidy assigned through a competitive procedure: those firms who apply are ranked according to a composite index of several criteria and funds are disbursed beginning from the top scoring firms going down the ranking, until the funds are exhausted. The point of exhaustion becomes the threshold, which cannot coincide with any other cut-off point, being created ad hoc. The firms cannot manipulate their score. Moreover, in a voluntary and competitive procedure, compliance is likely to be (nearly) universal: all those who applied take the subsidy if they score high enough to qualify.

A difference worth noting between the policies of Fig. 1 and of Fig. 2 is that the first is *compensatory* in nature (it targets subjects with *lower* values for the outcome variable), while the second intervention *rewards excellence,* the best cases are those who qualify. This implies that a naïve comparison between subsidized and unsubsidized firms would lead to an overestimate of the impact, while before we had the opposite problem: when the goal of the policy is to compensate for a disadvantage, a mechanical comparison of participants and non participants tends to underestimate impact.

Fig.2 Comparison around a threshold (for an intervention that rewards excellence)



A problem in many applications of RDD is that the number of observations around the threshold is too small for the test to have enough power (i.e., be able to statistically detect an effect if indeed an effect is there). The only way to overcome the problem of small sample size is to widen the band around the threshold—that is, include individuals that are further away from the threshold. This creates a trade-off between precision and bias: the precision of the impact estimate increases as the band widens, while also the bias increases as the band widens. If one opens up the band completely, one obtains a very precise and totally biased estimate.

The terms of the trade-off can be altered if one is willing to model the relationship between outcome and selection variable by using a regression model. The two lines in Fig.2 represent the linear regression of the outcome on the score assigned by the selection committee, separately for beneficiaries and non-beneficiaries. The estimate of the impact is given by the vertical distance between the two regression lines. The two lines tend to converge for lower values of the score, which has interesting policy implications—that is, that low scoring applicants would gain less, were they admitted into the programme, than those admitted.

It must be stressed that this rather interesting conclusion is quite fragile, being based on extrapolations. The two dashed portions of the regression lines represent the two counterfactuals, but they are mere linear projections of the solid portions, which by contrast are estimated with actual data. For example, we do not observe how beneficiaries would have performed had they not received the subsidy: we can only project the line representing the relationship between performance and score estimated for non-beneficiaries.

The only point where the impact estimate is not based on risky projections in at the threshold: for those scoring around 75 points, we compare estimates based on actual data. We are, a bit disappointingly, back to where we started: we adopt a regression model to use all the data we have, but the only estimate we trust is that based on the discontinuity in treatment around the threshold, which remains the true source of identification.

The main steps involved

The illustration of the main steps involved is based on the application of the method to an evaluation of the impact of R&D subsidies in France. The evaluation is by Nicolas Serrano-Velarde of the European University Institute, "The Financing Structure of Corporate R&D - Evidence from Regression Discontinuity Design", 2008.

Government intervention can only foster technological change if it supports projects that the private sector would not have implemented by itself. The study examines R&D subsidies given by the French ANVAR programme, responsible for R&D support to small and medium sized firms. ANVAR was created in 1979 to support R&D projects of small and medium sized firms through reimbursable aid. Every year ANVAR supports between 1.000 and 1.500 projects for a total budget of 250M Euros. Aid is paid on the basis of advancement of the project. Projects are selected on the basis of a bottom-up process by which firms propose their projects to the agency.

The empirical analysis combines the yearly R&D Survey from the Ministry of Research and the Financial Links Survey from INSEE. Pooling the data over the 1995-2004 period the database amounts to over 21.000 firm-year observations, while 2.312, approximately 11% of the firms in the sample, received financing from ANVAR.

Step 1. Specify the selection variable and the relevant threshold

The precondition for the applicability of RDD is the existence of a threshold that defines eligibility along a continuous variable. In order for a firm to be eligible for the ANVAR programme it has to be independent from a large business group (henceforth referred to as LBG). Independence is defined with respect to the firms' ownership structure, which becomes the selection variable. According to French law a firm is independent if less than 25% of its capital is owned by a LBG. Thus 25% becomes the eligibility threshold. A firm owned at 26% by a LBG will be considered ineligible in this setting.

Step 2. Choose the interval around the threshold

The author restricts the sample to firms which have $0\% < X < 50\%$ ownership by a LBG. Consequently, in the RDD sample a firm is marginally-eligible whenever it has positive ownership by a LBG but its share does not exceed 25%. Secondly a firm is marginally-ineligible whenever it has between 25% and 50% of its capital owned by a LBG. Excluded are also agricultural cooperatives or public related firms which escape firm categorization. The 1995-2004 full sample consists of 21.087 observations, whereas the 1995-2004 RDD sample consists of 566 firm observations which have $0\% < X < 50\%$ ownership by a LBG. Of these, one third (186) are ineligible, while two-third are eligible (380): of the latter, 294 are not treated while 86 are treated—that is, receive a R&D subsidy.

The study further distinguishes between four bandwidths around the threshold: Large ($0\% < X < 50\%$), Intermediate ($5\% < X < 45\%$), Small ($10\% < X < 40\%$), Very Small ($15\% < X < 35\%$). The smaller the bandwidth, the more likely are the conditions of a quasi-experiment. However one has to keep in mind the trade-off between length of the bandwidth and number of observations. For example, there are only 189 observations in the $15\% < X < 35\%$, including ineligible and eligible non-participants.

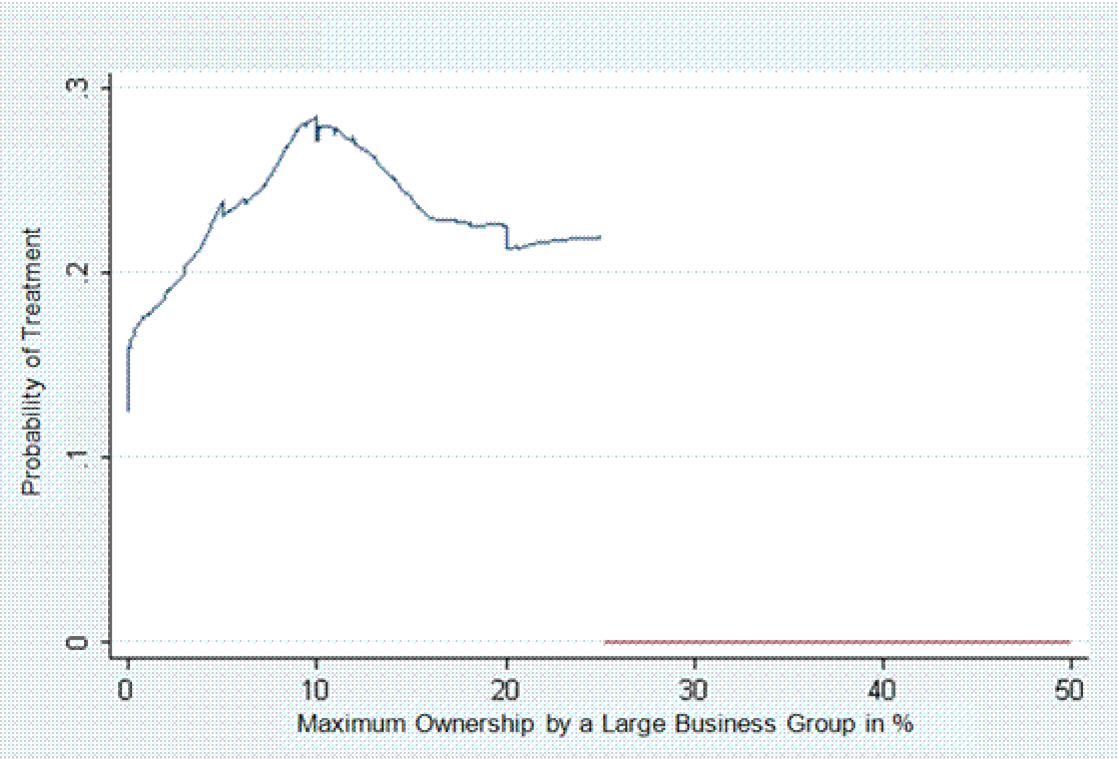
Step 3. Define the outcome variables

Private investment in R&D is defined as the relevant outcome variable. The hypothesis the study makes is that the subsidy should have an effect at lower levels of the R&D distribution and no effect at higher levels of the R&D distribution. The analysis then decomposes R&D investment into its internally and externally financed components. Effect at lower quantiles of the distribution should be driven by increased internal financing whereas public financing should simply substitute external financing at higher quantiles.

Step 4. An intermediate case between Fuzzy RDD and Sharp RDD

Ineligible firms in this setting have a zero probability of receiving the subsidy, whereas eligible firms have a positive assignment to treatment less than one, in the sense that only a small fraction of eligible firms takes up the subsidy. This creates a mixture of Fuzzy design (only 22% of eligible receive a R&D subsidy) and a Sharp design, as shown in Figure 3. Battistin and Rettore (2008) show that the conditions required to achieve identification in this setting are the same as in the sharp design. They show formally that, thanks to the discontinuity, eligible non-treated and ineligible firms are valid counterfactuals for supported firms, no matter how these supported firms self-select into the programme.

Fig.3 Probability of receiving the subsidy as a function of LGB ownership



Source: Serrano-Velarde, "The Financing Structure of Corporate R&D - Evidence from Regression Discontinuity Design", 2008.

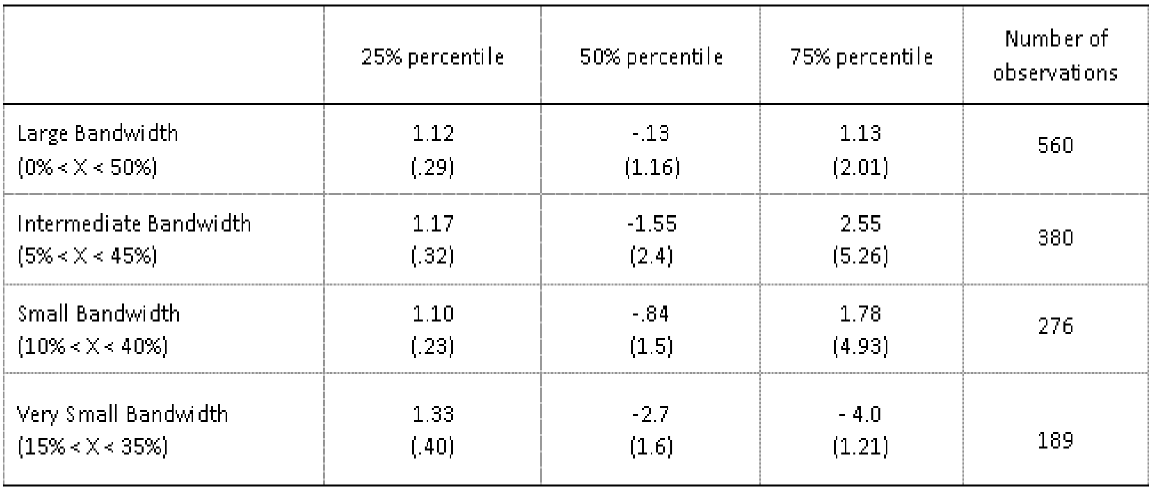
The line in the graph is obtained by a non-parametric regression technique, defined as "locally weighted smoothing regression" separately above and below the threshold of eligibility. A jump in the plot shows the effect of the threshold on the probability of receiving financing from ANVAR. It shows that there is a substantial effect of the eligibility threshold on the probability that a firm receives the subsidy.

Step 5. Estimating the effect of the subsidy

The author estimates a particular type of regression model, defined as quantile regression. The purpose is to go beyond the effect of the subsidy on the R&D expenditure of the average firms, estimating the effect on firms that are located at the 25%, 50% and 75% percentile of the R&D expenditure.

For all bandwidths considered the author finds a statistically significant positive effect of the R&D subsidy on private R&D investment for firms at the lowest quartile of the private R&D investment distribution. For firms with relatively smaller R&D budgets, an additional Franc of subsidy increases their own R&D investment by 1.1 Francs.

Tab.1 Quantile Regression Estimates



	25% percentile	50% percentile	75% percentile	Number of observations
Large Bandwidth (0% < X < 50%)	1.12 (.29)	-.13 (1.16)	1.13 (2.01)	560
Intermediate Bandwidth (5% < X < 45%)	1.17 (.32)	-1.55 (2.4)	2.55 (5.26)	380
Small Bandwidth (10% < X < 40%)	1.10 (.23)	-.84 (1.5)	1.78 (4.93)	276
Very Small Bandwidth (15% < X < 35%)	1.33 (.40)	-2.7 (1.6)	- 4.0 (1.21)	189

The following is the interpretation offered by the author. In firms where R&D is just of marginal importance to the business strategy, managers will have fewer incentives to be informed about R&D projects and their quality. Their imperfect information about the quality of their R&D activities leads them to invest less in innovation. Firms whose business model depends crucially on their ability to innovate will not face these internal doubts: they are obliged to invest in R&D. Managers of these ”strong innovators” have an incentive to be perfectly informed about the quality and the activity of their R&D department.

Strengths and limitations of the approach

This design allows one to identify the programme’s causal effect without imposing arbitrary exclusion restrictions, assumptions on the selection process, functional forms, or distributional assumptions on errors. RDD may be the best alternative to randomized studies for evaluating programme effectiveness. The most crucial element of the RDD design is its use of a ‘cut-off’ score on a pre-test measure to determine assignment to intervention or control. A valuable feature of this technique is that the selection measure does not have to be the same as the outcome measure, thus maximizing the programme’s ability to use research-based practice guidelines, survey instruments and other tools to identify those individuals in greatest need of the programme intervention.

On the other hand, the design features two main limitations. Firstly, its feasibility is by definition confined to those instances in which selection takes place on an observable pre-intervention measure. As a matter of fact, this is not often the case. Secondly, even when the design is feasible it only identifies the mean impact at the threshold for selection. Which in the presence of heterogeneous impacts tells us nothing about the impact on units away from the threshold for selection. In this sense, we only identify a local mean impact of the treatment. To identify the mean impact on the broader population one can only resort to a non-experimental estimator whose consistency for the intended mean impact intrinsically depends on behavioural assumptions.

Selected References

Battistin E. Rettore E [2008] *'Ineligibles and eligible non participants as a double comparison group in a regression discontinuity design'*, «Journal of Econometrics», vol. 142, n. 2,pp. 611-850

Cook T. [2008], *Waiting for Life to Arrive: a History of the Regression-discontinuity Design in Psychology*, Statistics and Economics, in «Journal of Econometrics», vol. 142, n. 2, pp. 636-654.

Hahn J., Todd P., Van der Klaauw W. [2001], *Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design*, in «*Econometrica*», vol. 69, n. 1, pp. 201-209.

Serrano-Velarde N. [2008], *The Financing Structure of Corporate R&D - Evidence from Regression Discontinuity Design*, Working Paper, European University Institute, Department of Economics, Firenze.

Thistlewaite D.L., Campbell D.T. [1960], *Regression-discontinuity Analysis: An Alternative to the Ex-post Facto Experiment*, in «Journal of Educational Psychology», n. 51, pp. 309-317.

Trochim W. [1984], *Research Design for Program Evaluation: The Regression-discontinuity Approach*, Beverly Hills, CA, Sage Publications.

Instrumental variables in details

The instrumental variables identification strategy

The fourth strategy is based on the idea of *involuntary variation* (in the official jargon *instrumental variables):* those situations in which the receipt of treatment is partially determined by an extraneous factor. As it will be apparent in the specific chapter, this identification strategy is notably more complex. The point of departure is that the structural effect of interest E cannot be recovered with any strategy based on the adjustment of S. There are no ways of forcing $S_{T\text{-}NT}$ to go to zero

$$\Delta_{T\text{-}NT} = E + S_{T\text{-}NT}$$

However the existence of the extraneous factor Z, which influences participation (to keep things simple we assume this to be binary), allows a way around the problem. One actually needs two identifying assumptions. The first is that the extraneous factor has an influence on T, in the sense that those with Z=1 participate in the policy with higher probability than those with Z=0. Thus we can write the effect of Z on T as

$$\Delta T_{Z\text{-}NZ} = T_Z$$

The second assumption is that the true effect of Z on the outcome can be recovered without any bias. This can be written as:

$$\Delta_{Z\text{-}NZ} = E_Z$$

Thus Z induces two effects: one on the outcome, one on participation. Neither effect is of much interest from a policy perspective, we are interested in the effect E of participation. It can be shown that E can be obtained by the ratio of the two effects of Z:

$$E = E_Z / T_Z$$

The proof of this result requires some algebra, while it is difficult to convey it intuitively

Description and purposes of the tool

This method is relevant when the exposure to a policy is not determined only by the decisions of the individuals involved, but also, to a significant degree, by events and processes outside their control. This "involuntary variation" in the exposure to a policy (to mimic the acronym of the name by which this method is known in econometrics, "instrumental variables") allows a rather ingenious way to eliminate selection bias. Others use the term *natural experiments*: Angrist and Krueger (2001) define natural experiments as those situations "*where the forces of nature or government policy have conspired to produce an environment somewhat akin to a randomized experiment*."

Whether they are called "involuntary variation", "instrumental variables", or "natural experiments", this approach has two essential ingredients: it requires that the exposure to the policy is to a certain degree determined by an "external force"; and that this external force does not affect the outcome of the policy directly, but only indirectly, through its influence on the exposure. If these conditions are met, the IV method produces credible estimates of the impact of the policy, although these estimates may be relevant only for the subgroup whose behaviour was changed by the external force.

An intuitive example

Take a programme to support private R&D research projects. Let us suppose that eligibility is restricted to firms in regions with low population density—a dimension which is not correlated with technological prowess and capacity. So, some "lucky" firms have access to the subsidy, others do not, but the propensity to conduct R&D research is not affected directly by the population density of the area. This is the crucial "identifying" assumption.

If all eligible firms took advantage of the subsidy, this would be essentially a situation similar to randomization. One would compare average R&D expenditure between the two groups and obtain an estimate of the impact of the subsidy. However, not all eligible firms take advantage of the subsidy: they self-select themselves according to their expected return from conducting R&D projects. Thus, a comparison of R&D expenditures of firms that apply for the subsidy and those who do not clearly would *overestimate* the effect of the subsidy, because of positive selection bias. It is even possible that most of the subsidised firms would have invested the same amount without a subsidy. If this were the case, the impact of the subsidy would be very low, possibly zero.

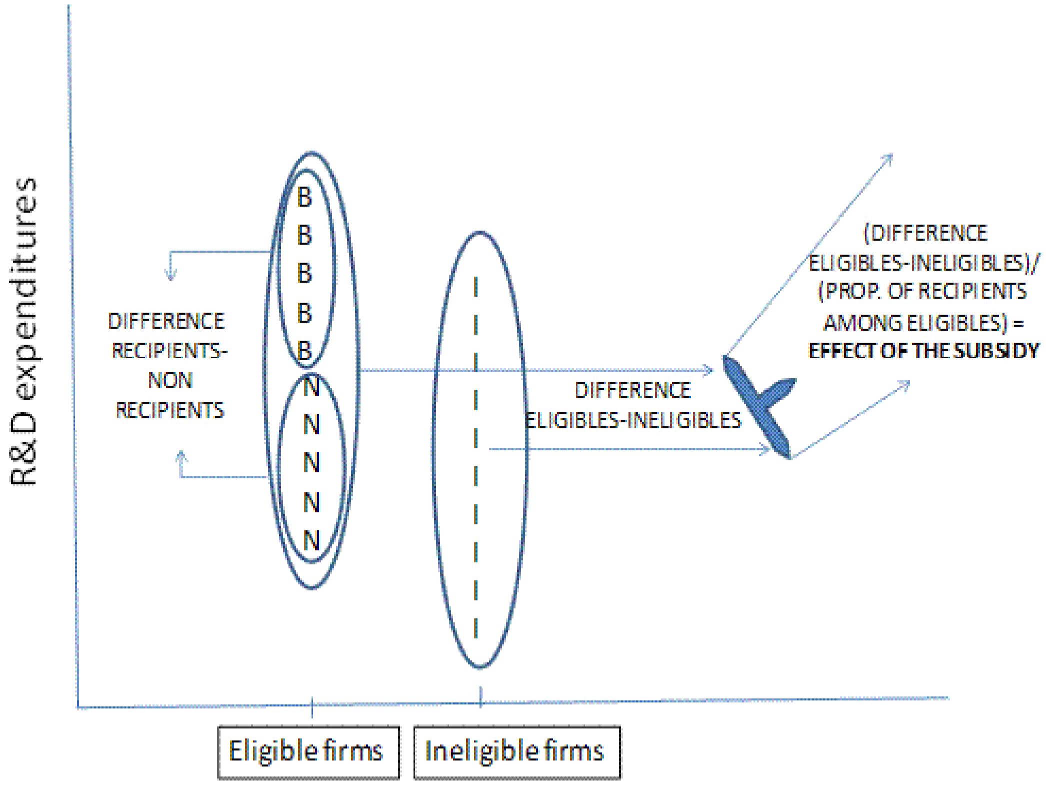
On the other hand, comparing the R&D expenditures of firms, eligible vs ineligible, only reveals the *effect of eligibility*, not the *effect of the subsidy*: we are generally more interested in the latter than in the former. Fortunately, there is a way to obtain a correct estimate of the effect of the subsidy: it is simply the effect of eligibility "scaled up" by the fraction of the eligible firms that take advantage of the subsidy. In practice, in this case one divides a difference by a take-up rate.

The intuition is that the difference in R&D expenditures between those eligible and those ineligible is accounted for by the fraction of eligible firms that actually receive the subsidy. For a given difference in R&D expenditures between eligible and ineligible, the smaller the fraction that receives the subsidy, the larger must be the true *effect of the subsidy*.

The effect of the subsidy is in fact obtained by "scaling up" the difference between eligible and ineligible firms—that is, by dividing by the fraction of those eligible who actually receive the subsidy. The rationale is simple: since the subsidy has no effect on the firms not receiving it, the whole difference between eligibles and ineligibles must be ascribed to the fraction of participants.

Figure 1 illustrates this intuition graphically. The B's represent the eligible firms that receive the subsidy. The N's the eligible non recipients, and together with the B's they make up the eligible firms. The ineligible firms are all indicated with an I. Moving from left to right, we see first the difference recipient/non-recipient, then the difference eligible/ineligible. The estimate of the effect of the subsidy is obtained by scaling up the eligible/non-eligible difference with the fraction of firms receiving the subsidy among those eligible.

Figure 1: An intuitive representation of the IV method (Wald estimator)


R&D expenditures
B
B
B
B
B
N
N
N
N
N
I
I
I
I
I
I
I
I
I
I
DIFFERENCE RECIPIENTS-NON RECIPIENTS
DIFFERENCE ELIGIBLES-INELIGIBLES
(DIFFERENCE ELIGIBLES-INELIGIBLES)/ (PROP. OF RECIPIENTS AMONG ELIGIBLES) = EFFECT OF THE SUBSIDY
Eligible firms
Ineligible firms

Circumstances in which it is applied

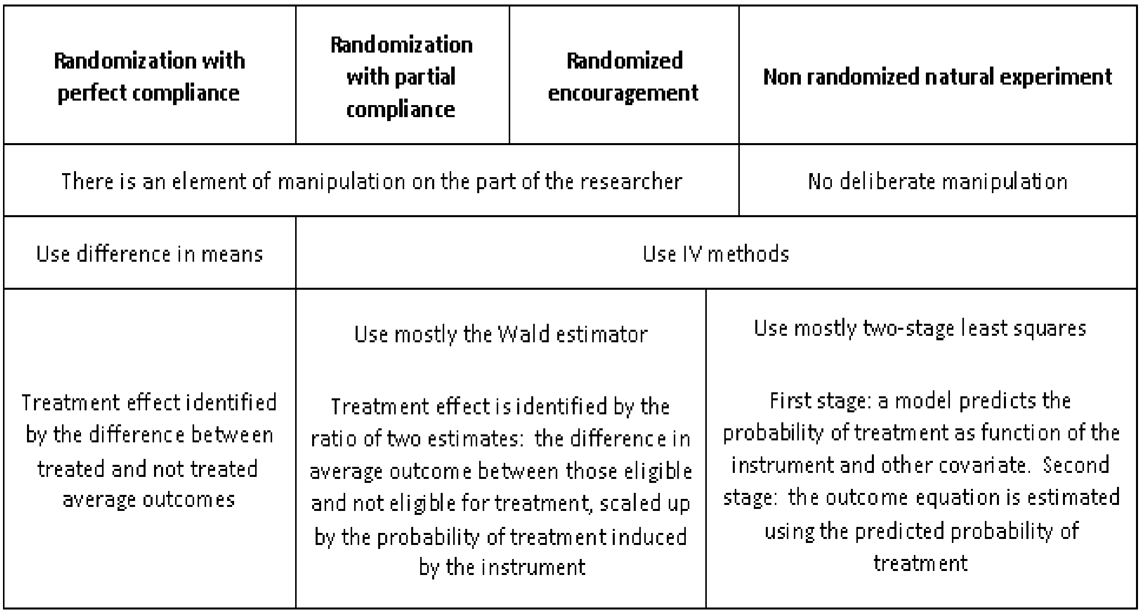
Half a century ago Norwegian economist Trygve Haavelmo (1944) emphatically spoke of "the stream of experiments that nature is steadily turning out from her own enormous laboratory, and which we merely watch as passive observers." Half a century has dampened this enthusiasm, and the stream of experiments has revealed itself not so steady: nevertheless, many examples of "involuntary variations in the exposure to a policy" can be found.

Examples include: idiosyncratic variations in administrative rules among adjacent jurisdictions; sudden changes in legislation, due the vagaries of the political process; geographical factors, such as the varying proximity of the client to the providers of services; unexpected shortfalls in funding for a programme; changes in administrative boundaries. These are all example of involuntary variation induced by a natural experiment. In the econometric literature these are called “instruments”, we refer to them also as the “extraneous force”.

Moreover, IV methods are applicable to all those situations in which the access to a programme was subject to randomization but the agents involved (the clients, the providers) did not fully comply with it, creating a situation in which programme access is co-determined by the agents’ preferences and by (what remains of) randomization. Finally, one should consider those situations in which an estimate of the policy impact is obtained by a special device: encouraging some people and not others, selecting the two groups randomly, to take part in the programme.

These different situations are summarized in Table 1. The first column represents randomization of a binary treatment with perfect compliance: no IV method is needed, impacts are given by simple differences in means. Next comes randomization with partial compliance: the treatment is still binary but some of those eligible (or ineligible) do not comply with the assigned status. Next comes a situation that also involves some lower degree of manipulation on the part of the researcher, who encourages some potential clients to participate in the programme but not others: the two groups are chosen at random. The last column represents the most classical situation, in which the involuntary variation occurs without intentional manipulation on the part of the evaluator.

Table 1: Situations in which the IV method can be applied



<table>
<tr><th>Randomization with perfect compliance</th><th>Randomization with partial compliance</th><th colspan="2">Randomized encouragement</th><th>Non randomized natural experiment</th></tr>
<tr><td colspan="4">There is an element of manipulation on the part of the researcher</td><td>No deliberate manipulation</td></tr>
<tr><td>Use difference in means</td><td colspan="4">Use IV methods</td></tr>
<tr><td>Treatment effect identified by the difference between treated and not treated average outcomes</td><td colspan="2">Use mostly the Wald estimator
Treatment effect is identified by the ratio of two estimates: the difference in average outcome between those eligible and not eligible for treatment, scaled up by the probability of treatment induced by the instrument</td><td colspan="2">Use mostly two-stage least squares
First stage: a model predicts the probability of treatment as function of the instrument and other covariate. Second stage: the outcome equation is estimated using the predicted probability of treatment</td></tr>
</table>

The main steps involved

We distinguish in this section two main approaches to IV estimation. One goes by the name of *Wald estimator* and is applicable to the simplest situations, where there is only one instrument with only two values: either the external force is on or is off. Moreover, we do not want to (or cannot) control for other characteristics. Technically, we speak of ”one binary instrument and no covariates”. This approach is limited in its applicability but very useful to understand the logic of IV estimation. It will require a higher degree of formalisation, exceeding that used so far to illustrate counterfactual methods. The reader not familiar with econometric reasoning and notation can skip the following pages and use the intuition gathered so far.

The second approach illustrated here ("two-stage least square") is more common in practical applications but considerably even more complex to explain: to make things more intuitive we will also give a concrete example of the two-stage procedure.

The Wald estimator

Step 1. Define the basic ingredients

The ingredients of the method are:

- Y is the outcome variable, R&D expenditures in the above example
- T is the binary indicator that is equal to 1 if the firm receives the subsidy, equal to zero otherwise
- U is what we do not observe about each firm that determines R&D expenditures
- Z is the "instrument"-- that is, the external force that *influences* T, but is *uncorrelated with* U (these are the two key assumptions for the IV method); it is equal to 1 if the firm is eligible, to zero in not eligible
- The symbol E() stands for "mean of" and the symbol E(|) represents the conditional mean.

Step 2. Define how the external force influences exposure

The next step is to model how the instrument influences participation—that is, the actual exposure to the policy. In the R&D example, when the subsidy is available, a positive fraction of firms apply for and receive the subsidy. In the areas with no funding, such fraction is close to zero (it might happen that some firms manage to get funding anyhow). Therefore we have that:

$$(1) \qquad \text{P(takes the subsidy} \mid \text{is eligible)} > \text{P(takes the subsidy} \mid \text{is not eligible)}$$

We can write with symbols what we just explained in words:

$$(2) \qquad P(T=1|Z=1) \; > \; P(T=1|Z=0)$$

Using a numerical example, the take-up rate of the subsidy P(T=1|Z=1) could be 0.75, while the degree of "cross-over" P(T=1|Z=0) (firms that are ineligible but manage to obtain the subsidy) is a much smaller 0.15. Thus the net effect of Z (the "instrument") on T ("participation") is equal to 0.75 – 0.15 = 0.60.

Step 3. Model the outcome of the policy

Let us write the outcome (R&D expenditures) as a function of the treatment (receipt of the subsidy) and of unobservable factors (the technological prowess of the firm).

$$(3) \qquad Y = T^{*}\delta + U$$

We know that the difference in outcome between beneficiaries and non beneficiaries of a policy, when participation is a choice, is the sum of the true effect and selection bias. Recall the expression we used in the chapter on the logic of counterfactual methods.

$$\Delta_{T\text{-}NT} \; = \; E \; + \; S_{T\text{-}NT}$$

Here we use equation (3) to show this more formally. By conditioning on T=1 first and on T=0 next and taking the difference, we obtain:

$$(4) \qquad E(Y|T{=}1) - E(Y|T{=}0) \;=\; \delta \;+\; [E(U|T{=}1) - E(U|T{=}0)]$$

The left-hand side of (4) is $\Delta_{T\text{-}NT}$, δ is the effect of the policy (E) and the term in brackets is the selection bias $S_{T\text{-}NT}$.

To continue with the numerical example, let us suppose the observed difference E(Y|T=1) – E(Y|T=0) is 50,000 euro. But we are interested in the value of δ, which could be any number, including zero. To do so we must use the "instrument".

Step 4. Using the instrument

Location in a low density area is uncorrelated with the technological prowess of the firm. Formally, the instrument Z is not correlated with the unobservable U. This is the crucial assumption of the IV method. It "excludes" any direct effect of Z on Y. It says, formally, that

$$(5) \qquad E(U|Z{=}1) - E(U|Z{=}0) = 0$$

This condition cannot be tested since U is not observable; it is an "identifying assumption". The reasoning continues as follows, requiring some more algebra. Let us partition the population of firms according to the value of Z.

Computing the means for those with Z=1

$$E(Y|Z{=}1) = \quad \delta{*}E(T|Z{=}1) \;+\; E(U|Z{=}1)$$

And for those with Z=0

$$E(Y|Z{=}0) = \quad \delta{*}E(T|Z{=}0) \;+\; E(U|Z{=}0)$$

The difference between the two equations is:

$$(6) \qquad E(Y|Z{=}1) \;-\; E(Y|Z{=}0) \;=\; \delta{*}[E(T|Z{=}1) - E(T|Z{=}0)] \;+\; [E(U|Z{=}1) - E(U|Z{=}0)]$$

The last term in brackets is equal to zero, as a consequence of the exclusion restriction. Thus (5) simplifies to:

$$(7) \qquad E(Y|Z{=}1) - E(Y|Z{=}0) = \delta{*}[E(T|Z{=}1) - E(T|Z{=}0)]$$

In addition, let us remember that the expected value of a binary variable is the same as the probability that the variable is equal to 1. Therefore:

$$(8) \qquad E(T|Z{=}1) \;=\; P(T{=}1|Z{=}1) \quad \text{and} \quad E(T|Z{=}0) \;=\; P(T{=}1|Z{=}0)$$

From (7) we obtain:

$$\delta \;=\; \frac{E(Y|Z{=}1) - E(Y|Z{=}0)}{P(T{=}1|Z{=}1) - P(T{=}1|Z{=}0)}$$

This is the Wald estimator. Both numerator and denominator can be estimated with observed data. If we estimate the difference E(Y|Z=1) – E(Y|Z=0) = 20,000 euro, the effect of the policy on R&D expenditure will be

$$\delta = \frac{20{,}000}{0.60} = 33{,}333$$

The two-stage procedure

The Wald estimator can be applied only to binary instruments and it does not allow the use of control variables. To overcome these limitations, a more complex procedure is often used, based on two stage regression procedure. To illustrate it, we will use a real example, taken from the study conducted by Elias Einiö (2009). The study recognizes the fact that "*the challenge in evaluating R&D support programs arises from the fact that subsidies are typically not randomly assigned, and as a result groups of supported and unsupported firms are not directly comparable*."

To overcome this problem the identification of the effect of programme participation is based on geographic variation in potentially available R&D-support funding arising from the allocation of European Regional Development Funds (ERDF) in Finland. These differences in allocation induce variation in the probability of programme participation, which facilitates the identification of the causal effect of programme participation on R&D effort.

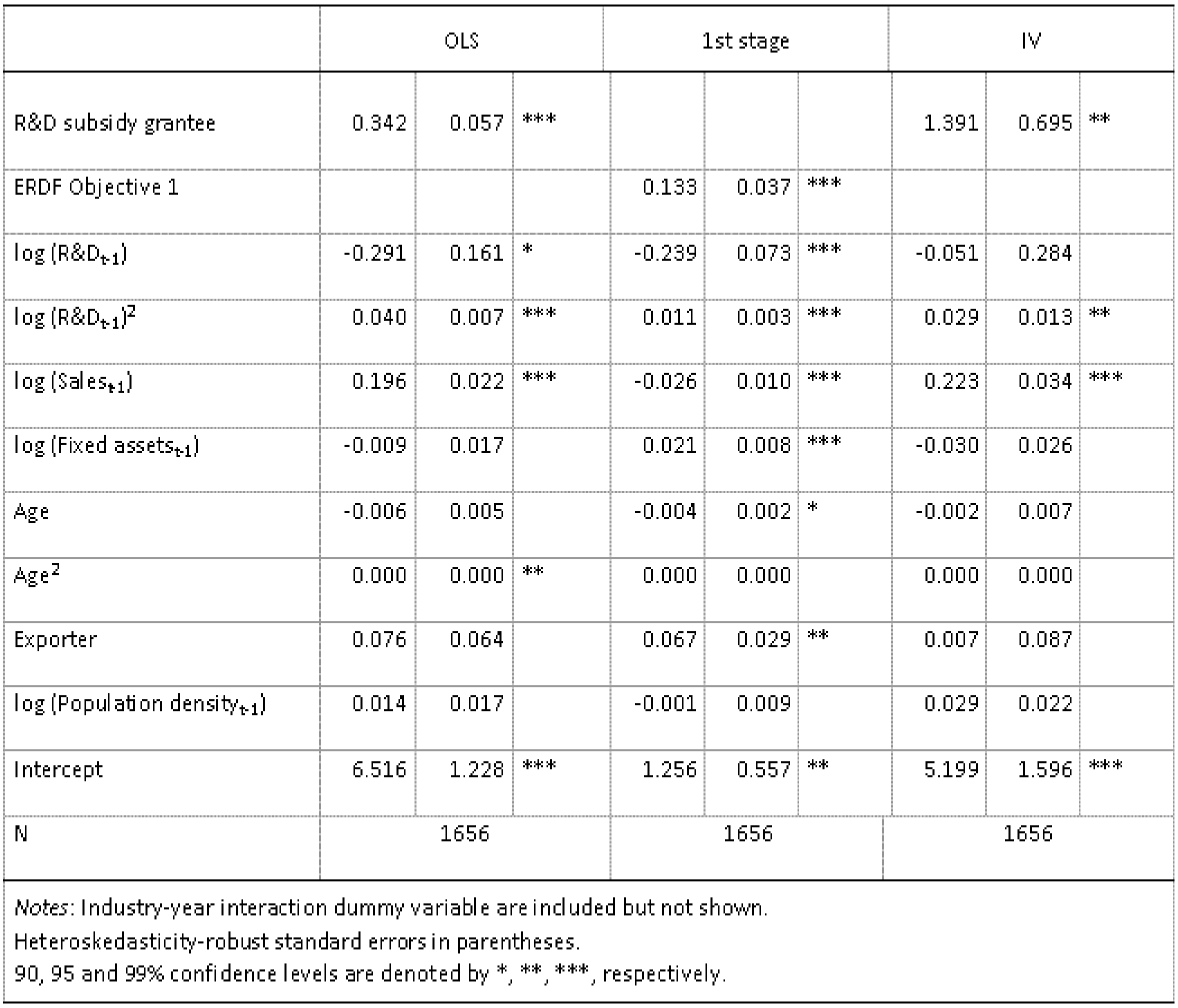
The advantage of the approach is that it is based on explicit differences in public policies with well-defined, publicly stated allocation criteria. The regional provision of ERDF in Finland is especially suitable for programme evaluation purposes because there are regions receiving the highest levels of European Union regional development aid because of low population density (rather than because of low levels of R&D investment or poor economic performance).

The average change in R&D expenditure among the supported firms from the year before they accessed the programme to the year after was €158,573, whereas among the unsupported group the corresponding change was €36,794. A simple before-after estimate suggests that the average effect of the programme was €158,573 - €36,794 = €121,779. However, this naive estimate is likely to suffer from selection bias because the support was not assigned randomly, and it is unlikely that the treatment and control groups are directly comparable.

The author turns to a OLS regression model, with the logarithm of R&D expenditure in year t + 1 as a dependent variable: the model controls for a vector of pre-treatment firm characteristics, including the log of sales and fixed assets in year t - 1 and a second-order polynomial of logged R&D in year t - 1 to control for permanent differences in the levels of R&D expenditure.

The results are shown in the OLS column of Table 2. The OLS estimate of 0.342 suggests that the support programme had a positive effect on R&D expenditure. However, if selection into the programme is affected by unobservable firm characteristics that also affect the R&D investment decision, the OLS estimate is likely to be biased. To overcome this problem, the author turns to an IV two-stage procedure, which allows the use of control variables.

Table 2: OLS and IV estimates Einiö's study on the effect of government subsidies on private R&D



	OLS			1st stage			IV		
R&D subsidy grantee	0.342	0.057	***				1.391	0.695	**
ERDF Objective 1				0.133	0.037	***			
log (R&D$_{t-1}$)	-0.291	0.161	*	-0.239	0.073	***	-0.051	0.284	
log (R&D$_{t-1}$)^2	0.040	0.007	***	0.011	0.003	***	0.029	0.013	**
log (Sales$_{t-1}$)	0.196	0.022	***	-0.026	0.010	***	0.223	0.034	***
log (Fixed assets$_{t-1}$)	-0.009	0.017		0.021	0.008	***	-0.030	0.026	
Age	-0.006	0.005		-0.004	0.002	*	-0.002	0.007	
Age^2	0.000	0.000	**	0.000	0.000		0.000	0.000	
Exporter	0.076	0.064		0.067	0.029	**	0.007	0.087	
log (Population density$_{t-1}$)	0.014	0.017		-0.001	0.009		0.029	0.022	
Intercept	6.516	1.228	***	1.256	0.557	**	5.199	1.596	***
N	1656			1656			1656		

Notes: Industry-year interaction dummy variable are included but not shown.
Heteroskedasticity-robust standard errors in parentheses.
90, 95 and 99% confidence levels are denoted by *, **, ***, respectively.

Step 1. The first stage of the IV method

In the first stage, a model is estimated of this form

$$T = \alpha + \beta Z + \gamma X + \varepsilon$$

X are other variables that we do observe and can serve as controls. The variable Z is equal to 1 when the firm is located in ERDF regions with low population density, zero otherwise. This represents the "instrument". The author motivates the choice on the instrument: "*We argue that differences in R&D support funding across the ERDF 1 border produce exogenous variation in programme participation: ERDF 1 eligibility is based on the criterion that the population density in the region is "no more than 8 persons per square kilometre" rather than on direct performance measures of the local economy.*"

The first-stage estimate[1] (second column of Table 2) for the "ERDF Objective 1" variable shows that the probability of programme participation is 0.133 points higher in the ERDF Objective 1 region, indicating that regional differences in available R&D support funding induce substantial differences in the probability of receiving support. If such difference is indeed induced at least in part by an external force - the ERDF eligibility rule—the difference in the outcome associated to such differential will be a measure of the true effect of the subsidy, purged of any component due to self-selection. Between recipients and non recipients there is a full difference of 1 in treatment, but most of it is due to self-selection: of that 1, only 0.133 worth of treatment is truly due to the external force.

If we were using a Wald estimator, we would compute the difference in R&D expenditure between eligible and ineligible firms and divide it by 0.133, thus “blowing it up” to obtain the effect of the subsidy. For example, if the difference in R&D expenditure between eligible and ineligible firms were €20,000 on average, the effect of receiving a subsidy would be on average €150,000. The study does not give the possibility of making these calculations, because the results are based on a logarithmic model, whose estimates represent percentage changes, not absolute differences.

Step 2. The second stage of the IV method

The second stage consists of estimating a model like

$$\log(\mathrm{R\&D}) = \theta + \delta\hat{T} + \varphi X + \upsilon$$

The logarithm of R&D expenditures is function of observable control variables (the same used in the first stage equation) and of $\hat{T}$, the predicted probability of participation from the first stage. From Table 2 we learn that the IV estimate of the causal effect of receiving a subsidy of programme participation is 1.391, which is significant at the 95 percent confidence level. The exponentiated value of the coefficient is exp(1.319) = 4.01, which brings the author to the following conclusion: “*This suggests that as a result of the programme the R&D expenditure among the supported firms was four times larger than it would have been in the absence of the support. Furthermore, in the extreme case of maximum subsidy compensation of 50 percent of the total post-treatment R&D costs, this result suggests that one subsidy euro induced at least 1.5 euro of additional company R&D*”

Step 3. Taking into account heterogeneity of impacts

A well-known result in the programme evaluation/econometrics literature is that, when the effects of the treatment are heterogeneous the IV estimate of the treatment parameter is the local average treatment effect (LATE), i.e. the average effect of the treatment among the participants whose treatment status the instrument changes. In the context of this study, LATE is the effect of programme participation among the firms entering the programme as a result of higher level of funding in the ERDF Objective 1 region.

Thus, the possibility cannot be ruled out that the effect of the programme would be different among the projects that would have received private support even in the absence of the ERDF funding. The author quotes the results of a beneficiary survey according to which more than 32 percent of the supported projects would have been undertaken even without support in the period 1999-2003 (Tekes 2007).

[1] It should be noted that it is not necessary to use probit or logit to generate first-stage predicted values, and it may even do some harm. In two-stage least squares, consistency of the second-stage estimates does not turn on getting the first-stage functional form right. So using a linear regression for the first-stage estimates generates consistent second-stage estimates even with a dummy endogenous variable.

This figure may largely underestimate the actual proportion of projects that would have been implemented even without government assistance because grantees may feel that revealing that government support was not necessary for the completion of the project may reduce the prospects of receiving assistance in the future. According to the results of the survey, at least one third, and plausibly even more, of supported projects were a priori privately profitable. The programme has quadrupled R&D expenditure among firms entering it as a result of higher government R&D-support funding in their region. The IV approach identifies the effect of the programme only among firms that change their participation status as a result of the higher funding in their region. Thus, the results should not be interpreted as evidence of the aggregate effectiveness of the programme.

[1] It should be noted that it is not necessary to use probit or logit to generate first-stage predicted values, and it may even do some harm. In two-stage least squares, consistency of the second-stage estimates does not turn on getting the first-stage functional form right. So using a linear regression for the first-stage estimates generates consistent second-stage estimates even with a dummy endogenous variable.

Strengths and limitations of the approach

The major weakness of the approach is that it can be difficult to find an instrument that is both relevant and exogenous. The assessment of instrument exogeneity can be highly subjective. Moreover, the IV method can be difficult to explain to those who are unfamiliar with it.

The major strength of the IV method is the fact of exploiting situations that are similar to a randomized experiment. Moreover, the use of researcher-generated instruments is growing and reflects the accelerating convergence of classical experimentation and observational research methods. The most important development is the use of instrumental variables in randomized experiments. Instrumental variables are useful in experiments when, either because of practical or ethical considerations, there is incomplete compliance in the treatment or control groups. In randomized evaluations of training programmes, for example, some treatment group members may decline training while some control group members may avail themselves of training through channels outside the experiment.

As in natural experiments, the instrument is used to exploit an exogenous source of variation—created by explicit random assignment in these cases—to estimate the effect of interest. Similarly, in medical trials, doctors may be willing randomly to offer, but not to impose, incentives that change behaviours like smoking or taking a new medication.

Progress in the application of instrumental variables methods depends mostly on the gritty work of finding or creating plausible experiments that can be used to measure important economic relationships. Here the challenges are not primarily technical in the sense of requiring new theorems or estimators. Rather, progress comes from detailed institutional knowledge and the careful investigation and quantification of the forces at work in a particular setting.

Selected References

Angrist J.D., Krueger A [2001] "Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments", *The Journal of Economic Perspectives*, Vol. 15, No. 4, pp. 69-85

Angrist J.D., Pischke J,S. [2008], *Mostly Harmless Econometrics,* Princeton University Press, NJ.

Einiö E. [2009], *The Effect of Government Subsidies on Private R&D: Evidence from Geographic Variation in Support Program Funding,* Discussion Paper n. 263 / May 2009, Helsinki Center of Economic Research, University of Helsinki.

Haavelmo T. [1944], *The Probability Approach in Econometrics*, in «Econometrica», Vol. 12, Supplement (July 1944), pp. iii-115.

9. Interviews

Description of the technique

Usually individual interviews consist of an in-depth conversation with an individual, conducted by trained staff. The purpose is usually to collect specific information related to the individual.

The interview technique is used to gather qualitative information and the opinions of those persons affected by a particular programme or project, its context, implementation and results. Several forms of interview can be distinguished, each of which fulfils a different purpose: the informal conversation interview; the semi-structured, guide-based interview; and the structured interview (the most rigid form). The following is based on a semi-structured interview, the form that is used most frequently in the evaluation of Structural Funds.

The purpose of the technique

Individual interviews, as the term suggestions, are a means of obtaining information through direct questioning.

In-depth interviews can help to:

Provide feedback on all aspects of a Programmes inputs, activities, outputs, outcomes and impacts. This type of survey is a way of learning about and examining the views of the actors (beneficiaries and other stakeholders) on a project or programme, e.g. how far the project or programme meets needs, or its results as compared to expectations. Interviews are also used in social science as a tool for investigating users' reasoning.

Provide a history of behaviour. When conducted more than once or when conducted with someone who has been involved in programmes for a long period of time, interviews can show if any change has occurred over time. The individual interview is an exploratory technique that serves to obtain relevant information on the reasoning, conceptions and representations of the persons questioned on a project or programme. Apart from subjective representations, it also serves to gather information on individual practices. It is particularly valuable in exploring the ways in which an intervention has been implemented and for identifying good practice.

Highlight individual versus group concerns. Topics that may not arise in a group situation can be addressed in individual interviews.

Reveal divergent experiences and "outlier" attitudes. Groups often do not allow you to see that experiences may vary person to person.

Provide a shortcut to community norms. Interviewing key community leaders (favourite teachers, police officers) can give a fast overview of a community and its needs and concerns.

Develop other research methods. Some evaluators use in-depth interviews to obtain information that they can then use to develop quantitative surveys once they have a better insight about what is occurring with a programme and what the key issues are. Others find that interviews give them all the information they need without conducting a later survey. Results from an interview can be used to generate focus group questions or help form questions for a survey. In-depth interviews can be different from focus groups in several ways:

In-depth interviews are also often to get preliminary ideas from stakeholders.

It is often used as part of a formative evaluation that is designed to test a theory of action and/or to provide guidance about fine-tuning a policy and programme.

It is helpful in providing a summative evaluation of a programme which is intended to achieve changes in actors' behaviour or perceptions (such as technology transfer or training programmes) as opposed to the more tangible, 'harder' outputs (such as creating new jobs or constructing new facilities).

Circumstances in which it is applied

Every evaluation usually involves individual interviews, such as talking to programme managers and stakeholders, and often beneficiaries and participants. One cannot say that interviews are always appropriate or inappropriate, but rather they are preferable for some types of evaluation, under some conditions. Indeed in some cases the choice of whether to conduct a personal interview will be made for extraneous reasons such as costs or pressure of time. For the circumstances and benefits see below.

The interview is used in an exploratory context, in other words, when one does not have a priori hypotheses or adequate knowledge on a project or a target public to make a questionnaire survey possible.

It is also a relevant technique when the stakeholders involved in the evaluated programme or project are too small in number to be the subject of a statistically representative survey.

Interview techniques are used extensively in the evaluation of structural interventions and, in particular, for the evaluation of programmes. Interviews with programme managers and beneficiaries remain one of the most commonly used methods in the intermediate evaluation of structural programmes. The principle consists of selecting several beneficiaries or managers, depending on the characteristics of the action implemented and the public concerned.

Used as a method for analysis, the interview is one of the only techniques which makes it possible to make a brief overview of the programmes. Very loosely structured interviews - informal conversations with the managers and other persons concerned by the programme - may be used to identify those parts of the programme which need to be considered in more depth. Interviews with policy makers may also form an important part of the early stage of focusing an evaluation and determining the key outputs which policy makers require. By conducting individual interviews with the key stakeholders and establishing their priorities, the evaluator also gets an insight into what people see as priority topics to be addressed by the evaluation.

The interviews may also prove relevant and provide useful information in the framework of programmes of a social nature, where the beneficiaries often lack motivation for filling in the questionnaires, such as measures aimed at giving jobs to the long-term unemployed who have been on a training course. Furthermore, key stakeholders often provide information on other stakeholders that they feel should be consulted, and how best to access and engage them.

The main steps involved

In order to conduct an interview well, the following steps have to be adhered to.

Step 1. Selection of interviewees

The samples needed for carrying out interviews are smaller than for questionnaire surveys. The information obtained is validated by the context and not by the probability of occurrence peculiar to questionnaires. The number of interviews depends on the subject of the study, on the variety of reactions to the subject, and on the resources available (in general 20 to 60 interviews). The selected sample is most often based on the selection of components said to be characteristic of the population (diversified sample). The sample may be selected either by direct access (e.g. via administrative lists, memberships of various relvevant bodies / organisations) or through the intervention of a third party, which allows for more targeted selection but presents risks of distortion.

Step 2. Planning the interview

Planning the interview includes the drawing up of an interview guide. This consists of specifying the topics that the interviewer wants to address. It is not essential to follow the interview guide in any precise order. This guide is more a sort of checklist enabling the interviewer to check that he or she is dealing with the essential questions. The interviewer may modulate her/his intervention in relation to the interviewee, and formulate new questions. The first interviews may give rise to adjustments or amendments in the definition of the questions if the interviewees have a problem with them. In many cases, it is also helpful to gather basic information from interviewees in advance to save time during the interview itself. It may also be useful to provide the interviewee with a guide to the issues that are to be covered so that he or she can gather together any necessary factual information prior to the meeting. Certainly it is usually worth establishing a 'contract' with the interviewee including explaining the purpose of the interview, how long it will last, the level of confidentiality, the use made of findings etc. It is useful to understand prior to interview what the stakeholder's role is as interview schedules will vary according to what their relationship is with the programme.

Design and piloting of questions - this could range from a checklist to a semi-structured questionnaire.

Step 3. Selection and training of interviewers

If they are to be conducted properly, interviews require a high degree of professionalism in the interviewer. He or she must have skills in communicating, listening and note taking. To facilitate the smooth running of the interview and ensure that the interviewee feels at ease, it may be useful to ensure that there is social proximity between interviewer and interviewee. The least structured interviews also require the interviewer to have substantial knowledge of the field.

Step 4. Course of the interview

An interviewer must have a "respectful" attitude vis--vis the interviewee and the information gathered, but also be able to convey a good understanding of the subject matter and context. The initial contact is very important, for it is the basis of communication. The interviewer must be careful not to influence the interviewee by approving or orienting her/his answers. The interview may be recorded, to ensure that the interviewee's assertions are not distorted and that the most important remarks are not omitted from the report. However, in some cases interviewees may feel less able to give full and frank answers on tape, and the costs of transcribing interviews should be taken in to account in designing the evaluation budget and timetable. The interview transcript or summary may be checked with the interviewees and it may be useful at this stage to follow up any unresolved questions. It may also be useful to thank interviewees by letter for their contributions.

Step 5. Analysis of results

This final phase consists of analysing the conversations, interpreting and comparing the information given by the interviewees, and finding common and divergent viewpoints so as to draw up a review of the evaluation. A summary focused on the coherence of each interview, or a transversal thematic analysis more suited to the search for models capable of explaining individual practices, may be drawn up.

In order to produce results and then discuss then, the interviewer establishes an analysis grid based on the reading of the interviews and of descriptive hypotheses. This grid is an explanatory tool applied to each interview report.

Once the research has been conducted and written up, all papers, cassette tapes, etc. should be kept secure. Ideally, only selected staff should have access to the data, such as interviewers, supervisors, staff who record or verify the data, and data analysts.

Strengths and limitations of the technique

Individual interviews are probably crucial to the Structural Funds evaluations. To get a clear view on complex issues, in depth research and understanding of the complex issues is required, as opposed to a broad approach to the research. By using individual interviews, the views of individual respondents and the reasons for this opinion can be discovered, without any influence from other actors (as may be the case in a focus group situation). Individual interviews are especially suited for getting insight into process issues.

The advantage of this type of method is that it provides in-depth information on the values, facts and behaviour of the interviewees; it makes it possible to link up a group of elements, thus producing a relatively exhaustive study on a given subject. A well-conducted interview may provide insight into the mechanisms of implementation and the causal links peculiar to a programme, and help to identify success stories or obvious shortcomings. Indeed it can help to propose solutions and recommendations for taken the programme forward. As a result, the technique produces information that can easily be communicated in the form of boxed examples in a text. In addition to this, the sample can be controlled and the interviewer has confidence that their questions have been interpreted as intended.

In certain cases, individual interviews are preferred over other methods such as group discussions, for example when a subject is surrounded with strong social norms, or when a judgement is required. Interviews are also one of the best ways to engage low-literacy populations. Structured interviews can take the place of questionnaires for clients who may have difficulty filling out forms (i.e. problems).

Furthermore, individual interviews are often easier than group methods. The key benefit is the level of detail that can be obtained. In an interview evaluators have a chance to follow-up on questions and probe for meaning. It can be easier to discuss an issue in-depth with one person, than with a group. It also helps avoid the scheduling problems of trying to arrange meeting dates with large numbers).

However, there are drawbacks. When data is obtained through in-depth interviews, the sample size is usually smaller and does not use random methods to select the participants. Moreover, an individual interview takes into account situational and individual factors making it difficult to draw general conclusions. Individual interviews may allow for an exhaustive identification of effects and possible causes, but cannot be used to measure impacts or grade causes.

If it is to be effective, this practice requires a lot of time and the contribution of professionals. Specific skills are needed to plan, conduct and interpret an interview; inadequate skills of this nature will produce information of no value. Interviews are not always conducted with the professionalism required to produce effective results. Other dangers of this method that must be guarded against are subjectivity: despite the use of trained interviewers, there is still the risk of a lack of consistency between interviewers. The individual interview as a method thus suffers from subjectivity, and relies heavily on the integrity and intellectual honesty of the researcher, whose experiences cannot be replicated, by the very nature of the research. On the other hand, personal interviews also eliminate the chance of anonymity and the interviewer may influence the answers.

Furthermore, the benefit of these method depends on the knowledge of the interviewees and on their co-operation in answering the questions. Information may also be distorted due to the choice of interviewees.

Documentation can be tricky: field notes often contain too much confidential information for wider circulation: much has to be taken on trust in the reporting stages. Importantly evaluators should be respectful of privacy when discussing specific clients or respondents and avoid being drawn into critical conversations or providing confidential information regarding other interviews. The role is to fact find and gather useful information, not to inform the subjects of any opinions they may have. At no time should a research subject be referred to by name during staff meetings or debriefings.

Bibliography

Blanchet A., Gotman A. (1992) L'enqute et ses mthodes: l'entretien. Nathan Universit This book is entirely devoted to interview surveys, from an educative point of view.

Bryman, A., (2001) Social Research Methods, Oxford Publications

Hermatti, Minu (2002) Multi-stakeholder processes for governance and sustainability: beyond deadlock and conflict, London: Earthscan, 2002

Patton M.Q., (1987) How to use qualitative methods in Evaluation, Sage Publications voir dernière édition

Oppenheim, A. (1992) Questionnaire Design, Interviewing and Attitude Measurement

10. Models

For readers wishing to obtain information on modeling of the Structural Funds, the paper at the attached link is recommended:

http://ec.europa.eu/regional_policy/sources/docgener/work/2010_02_modelling.pdf

11. Multi-criteria analysis

Description of the technique

Multicriteria analysis appeared in the 1960s as a decision-making tool. It is used to make a comparative assessment of alternative projects or heterogeneous measures. With this technique, several criteria can be taken into account simultaneously in a complex situation. The method is designed to help decision-makers to integrate the different options, reflecting the opinions of the actors concerned, into a prospective or retrospective framework. Participation of the decision-makers in the process is a central part of the approach. The results are usually directed at providing operational advice or recommendations for future activities.

Multicriteria evaluation be organised with a view to producing a single synthetic conclusion at the end of the evaluation or, on the contrary, with a view to producing conclusions adapted to the preferences and priorities of several different partners. In the case of European Union socio-economic programmes, the different levels of partnership (European, national and regional) may be concerned. Each of these levels is legitimate in establishing its own priorities and expressing its own preferences between criteria.

Multicriteria analysis is similar to the techniques adopted in the field of organisational development or information systems management. It also resembles cost-benefit analysis although it does not reduce the disparate phenomena to a common unitary (monetary) base.

The purpose of the technique

The purpose of the tool is to structure and combine the different assessments to be taken into account in decision-making, whereby decision-making is made up of multiple choices and the treatment given to each of the choices condition the final decision to a large extent. Importantly, multicriteria analysis is used to highlight the reasoning and subjective convictions of the different stakeholders on each particular question. It is usually used to synthesise the opinions expressed, in order to determine the priority structures, to analyse conflictual situations, or to formulate recommendations or operational advice. The applications could include for example:

- Making recommendations on the reallocation of budgets, either while the programme is underway or during the preparation of the following programme. The main decisions in this respect are taken at the level of the intervention or priority. Interventions judged to be the least successful must be re-examined with a view to either reducing their budgets or re-organising them to enhance their effectiveness. Where relevant, recommendations can also be made to increase the budgets of those interventions ranked as being the best.
- Diffusion of good practice, by identifying the areas of success and the most effective interventions of the programme. Information on those measures judged as being the most successful (good practice) can be disseminated through a range of means, including the media, if the authorities running the programme wish to show the public how funds were spent.
- Publishing concrete examples of successful projects or interventions can also help to inform the managers of similar interventions financed elsewhere.

- Feedback on project selection methods. The choice of evaluation criteria, their precise definition and their weighting constitute a useful contribution to multicriteria analysis. This work makes it possible to formulate a clear, complete and coherent description of the intentions and priorities of the programme partners. It is then possible to use these results to spread clear messages to the managers of the interventions.
- Enhancing the project selection process. It is relatively easy to transfer criteria, scoring scales and weightings to the project selection system if this system is also organised on the basis of scoring-weighting. By basing the selection of projects on the same logic as the evaluation of measures, the chances of stimulating and funding projects which contribute effectively to the programme priorities are increased.

Multicriteria analysis is well suited to managing and evaluating structural programmes in partnership since the opinions of national and supranational members may be expressed jointly without losing any of their specificity or having to make too many concessions regarding their value scales. Multicriteria analysis was used in the evaluation of a regional development programme co-financed by the three European Structural Funds and the government of the Walloon region: in this case, a variation of the method was developed called "multicriteria-multijudge" analysis, which enabled each partner to construct her or his judgement based on the criteria and weights of her or his own choice.

Circumstances in which it is applied

Multicriteria analysis is a tool for comparison in which several points of view are taken into account, and therefore is particularly useful during the formulation of a judgment on complex problems. The analysis can be used with contradictory judgment criteria (for example, comparing jobs with the environment) or when a choice between the criteria is difficult.

In general, this technique is mainly used in ex ante evaluations of public projects and their variations (the layout of a highway, the construction of a new infrastructure, etc.). Less commonly however, multicriteria analysis is also applied to the intermediate or ex post evaluations of programmes. However, it probably has potential for wider use as a tool in intermediate and ex post evaluations as an aid for making a judgment. Within the framework of socio-economic development programmes, it concerns a judgment on the success of the different measures, for the purpose of drawing synthetic conclusions. This judgment takes into account the main relevant criteria for the steering group.

The main steps involved

The main steps involved in multicriteria analysis can be broken down into several phases described sequentially below. It is possible to repeat the phases and thus to make corrections.

Step 1. Definition of the projects or actions to be judged

This will involve an inventory of the planned or implemented actions, or the elements on which the comparative judgment will be made.

Step 2. Definition of judgment criteria

Particular attention must be given to the definition of criteria, in order to be as exhaustive as possible and to define the question properly. The criteria must reflect the preferences of the decision-makers or the different points of view, so as to summarise and group together diverse characteristic dimensions used to evaluate an action.

If the evaluation was intended to focus primarily on the relevance of the programme to the regional economy rather than the impacts, the multicriteria analysis would concentrate on the main strengths and weaknesses of the regional economy and the way in which the different priorities build on strengths or offset weaknesses.

The synergy between the impacts of the different interventions or priorities could also be considered, and if so 'synergy' would become a judgment criterion in its own right.

Unlike the number of interventions to be compared, which can be very large, the number of criteria must not exceed a reasonable limit. Experience has shown that the maximum number of criteria for an effective evaluation is eight criteria.

A key issue in multicriteria analysis is the involvement or not of the different actors in the definition of criteria and their weighting. If the evaluator is actively involved in the analysis, the credibility of the results is undermined. On the other hand, when the stakeholders of the evaluation participate in the definition of the criteria, each partner prolongs the discussion until at least one judgment criterion is found that places her or his favourite action in first position. Usually the commissioners of the evaluation will have final say in specifying the criteria.

Before continuing with the multicriteria analysis, the evaluation team must check whether the process will allow for interventions to be compared satisfactorily. In choosing the criteria, the team should already have ensured that they apply to as many interventions as possible. The majority of these must have produced impacts related to the majority of criteria (that is, the impact scoring matrix must not have too many neutral, absent or insignificant impacts). The example below shows a situation where the only scores which are not equal to zero are situated in the diagonal. This suggests that the interventions to be evaluated have nothing in common. Therefore the evaluation criteria are intervention-specific and multicriteria analysis cannot be performed.

Case of evaluation criteria that are too specific

Criterion	*Diversification*	*Employability*	*Environment*
Modalities for evaluating the criteria	*(impact rating between 0 and 10)*	*(impact rating between 0 and 10)*	*(impact rating between 0 and 10)*
Interventions			
investment aid	*7*	*0*	*0*
in-house training	*0*	*5*	*0*
industrial redeployment	*0*	*0*	*8*

Step 3. Analysis of the impacts of the actions

Once the interventions and criteria have been defined, a quantitative estimation or a qualitative description of the impact of each project, in terms of these criteria, must be made. For this purpose short statements describing the different levels of impact may be used ("impact descriptors").

Based on the judgment criteria and measures (or groups or parts of measures) to be evaluated, the evaluation team would usually construct a multicriteria evaluation matrix. This matrix is a table with as many columns as there are criteria and as many lines as there are interventions to be compared. Each cell represents the evaluation of one intervention for one criterion. Multicriteria analysis requires an evaluation of all the interventions for all the criteria (no cell must remain empty), but does not require that all the evaluations take the same form. As shown below, the technique can support a mix of quantitative criteria expressed by indicators, qualitative criteria expressed by descriptors, and intermediate criteria expressed by scores.

Combination of qualitative and quantitative criteria

Criterion	*Diversification*	*Employability*	*Environment*
Modalities for evaluating the criterion	*(% of assisted businesses active in growth sectors)*	*(impact descriptors)*	*(impact rating between 0 and 10)*
Interventions			
Investment aid	*72%*	*Neutral impact*	*3*
In-house training	*21%*	*Significant increase in the employability of trainees already qualified; neutral impact for the others*	*1*
Industrial redeployment	*52%*	*Neutral impact*	

Two main possibilities exist for the evaluation team, for comparing the merits of the different interventions using scoring:

- multicriteria analysis by compensation or
- multicriteria analysis based on outranking.

These methods are described below. Outranking does not always produce clear conclusions, whereas analysis based on compensation it is always conclusive. From a technical point of view, the compensation variant is also easier to implement. The most pragmatic way of designing the multicriteria evaluation matrix is for the evaluation team to design scoring scales to all the evaluation conclusions, whether quantitative or qualitative. The multicriteria evaluation matrix is then equivalent to the impact scoring matrix. Usually the compensation method is used unless members of the steering identify a problem which might justify the use of the veto system.

Compensation method

The compensation method is the best-known variant and consists of attributing a weight to each criterion and then of calculating a global score for each measure, in the form of a weighted arithmetic average of the scores attributed to that measure for the different criteria. This variant is called "compensatory" because the calculation of the weighted average makes it possible to compensate between criteria. For example, an intervention which had a very bad impact on the environment could still obtain a good global weighted score if its impact on employability were considered excellent.

Outranking method

The outranking variant is used where the criteria are not all considered commensurable, and therefore no global score can be produced. The analysis is based on multiple comparisons of the type: "does Intervention A outrank Intervention B from the point of view of the environment criterion?", "does Intervention A outrank Intervention B from the point of view of the employability criterion?", etc. These questions can be answered yes or no or be qualified, in which case the notions of a weak preference and a threshold criterion are introduced. The analysis makes all possible comparisons and presents a synthesis of the type: "Intervention A is at least as good as Intervention B, in relation to a majority of criteria (case of agreement), without being altogether too bad in relation to the other criteria (case of disagreement)".

The analysis could include protection against a favourable judgement for an intervention that would be disastrous from the point of view of the given criterion, by setting a 'veto threshold' for each criterion. The introduction of a veto threshold strongly differentiates the logic of outranking from the logic of compensation. If there were a veto threshold, a very bad impact on the environment would make it impossible to consider the measure good, even if its impact on employability were considered excellent.

Outranking has the advantage of reflecting the nature of relations between public institutions better, since there is often a correspondence between evaluation criteria and evaluation stakeholders. In cases where the steering group is extended to about ten partners, it is not unusual for participants to identify themselves strongly with the "environment" or "employment" criteria. In this situation the outranking variant will probably better reflect the collective process of formulating a judgement within the steering group.

Step 4. judgment of the effects of the actions in terms of each of the selected criteria

This involves evaluating the impacts. If the compensation methods is used the process involves allocating scores, and a simple analysis using a basic spreadsheet. For the outranking variant, the approach will differ according to the type of analysis, of which the most well-known are presented below.

Variants of multicriteria analysis using outranking

ELECTRE I - This variant functions with an agreement index and a disagreement index, presented in the form of scores. A disagreement threshold (a veto) is introduced for all the criteria. The outranking and veto thresholds are of the franc type. The software processes a situation in which the best measure(s) must chosen, for example a situation in which the aim is to identify best practice.

ELECTRE TRI - This variant serves to sort measures into different categories, for example, the most successful measures, measures which have no significant impact and intermediate measures.

ELECTRE II produces a ranking of measures, from the most successful to the least successful. Outranking and veto thresholds are of the franc type.

ELECTRE III also performs a classification, but introduces vague outranking relationships.

PROMETHEE uses only an index of agreement and introduces progressive outranking.

For more information, see the annexed bibliography: Vincke 1989.

The process could be based on quantitative data, or, undertaken more subjectively, by experts or the stakeholders of the evaluation themselves. In reality, the technique usually combines factual and objective elements concerning impacts, with the points of view and preferences of the main partners.

In collecting the views of the partners, the evaluation team usually uses individual interviews or focus group interviews with those people whose points of view are considered most relevant for judging the interventions (referred to as the 'assessors'). A popular option is to use the members of the evaluation steering group as assessors. Ideally, the steering group should be large enough to reflect the main points of view, but a group of six to ten assessors is probably optimal, and therefore they would tend to be a subset of the wider steering group.

The assessors' preferences are taken into account according using one of several methods:

Through direct expression in the form of a weighting attributed to each criterion. This can be done by means of a vote through the distribution of points. The discussion can also be conducted by means of several successive meetings.

Revealing preferences by classification of profiles. In this variant the assessors are presented with "profiles" of measures or projects described in such a way that they reveal preferences between criteria. The assessors have to choose one of these two profiles and, if possible, must state whether their preference is weak, average, strong or very strong. The exercise is repeated for all the pairs of profiles, and a software package is used to attribute a weight to each impact, expressed as a percentage so that the weightings add up to 100%.

Revealing preferences through the ranking of real projects. The choice offered to the assessors in the preceding variant could have the drawback of seeming artificial. To avoid this problem, it is preferable to ask the assessors to state their preferences between real projects.

Step 5. Aggregation of judgments

Usually a computer package is used to sort the actions in relation to each other. A single weighting system for criteria can be deduced, or the evaluation team and steering group can decide to establish average weightings, which has the effect of effacing different points of view among the assessors.

There are three different approaches to the aggregation of judgments:

- Personal judgments: the different judgment criteria are not synthesised in any way. Each of the addressees of the evaluation constructs her or his own personal judgment based on the analysis and uses it to argue her or his point of view.
- Assisting coalition: the different judgment criteria are ranked using a computer package. An action will be classified above another one if it has a better score for the majority of criteria (maximum number of allies) and if it has less 'eliminatory scores' compared to the other criteria (minimum number of opponents).
- Assisting compromise: a weighting of the criteria is proposed by the evaluator or negotiated by the addressees of the evaluation. The result is a classification of actions in terms of their weighted score.

In the most common application of the method at this stage, the evaluation team has all the elements it needs to calculate global weighted scores for the different measures. The results of each measure will have been evaluated in relation to the same criteria; all these evaluations will have been presented in the form of scores in an impact scoring matrix; there is a weighting system which expresses the average preferences of assessors for a particular criterion. The global score is calculated by multiplying each elementary score by its weighting and by adding the elementary weighted scores . Based on weighted average scores, the evaluation team can classify measures by order of contribution to the overall success of the programme.

The 'multi-judge' variant consists of maintaining the individual weightings of each assessor. In this case, the anonymity of the assessors must be respected when the weightings carried out individually by them are processed. However, if preferences among criteria show strongly divergent points, it is possible to establish several classifications of the measures. On the same impact scoring matrix, the evaluation team can apply different systems of weighting (a play of different weightings for each assessor). Differences between the weighted global scores and hence differences in ranking will result, since each measure can be considered a success from the point of view of one assessor and a failure from the point of view of another. It may then be interesting to present the weightings separately, for a particular category of assessor, for example, whether the assessors claimed to identify more with national or regional concerns.

The synthesised judgment on the effectiveness of measures is usually considered sound and impartial provided that:

- the evaluation criteria have been validated by the steering group;
- the conclusions on the impacts of each intervention, as well as the impact scoring matrix summarising them, and have been validated;
- the weighting coefficients for criteria, have been established with the assistance of the assessors and the agreement of the steering group.

Experience also shows that the partners are far more willing to accept the conclusions of the report if the evaluation team has recorded their opinions carefully and taken the trouble to take their preferences into account in presenting its conclusions. If, on the contrary, the evaluation team chooses and weights the criteria itself, without any interaction with its partners, the impartiality of the results will suffer and the multicriteria analysis will be less useful.

Strengths and limitations of the technique

As mentioned already, multicriteria analysis provides a framework in which all the actors can take part in decision-making and in problem solving. Through negotiation between stakeholders and explicit treatment of judgment criteria, the technique serves to give form to an unstructured reality. The strength of multicriteria analysis therefore, lies in the fact that it allows the values and individual opinions of several actors to be taken into consideration, and the processing of functional relations within a complex network, in a quantitative way.

The intervention of an expert, the margin of manoeuvre enjoyed by decision-makers and similarities with vote-based methods makes this a suitable tool for a partnership approach.

The technique is well suited to the way in which partnerships function in so far as it outlines areas of consensus in which the partners agree on the ranking of measures, and areas of dissension which reveal the interventions considered successful for some and unsuccessful for others. Experience has shown that consensual conclusions are generally in the majority. This can be explained by the fact that the different weightings apply to the same impact scoring matrix. Thus, a measure which has a low score for all the criteria will never have a high weighted global score, irrespective of the differences of priorities between partners. The different points of view of the partners cannot strongly contradict the conclusions resulting from empirical observation if these conclusions show that certain measures are really part of good practice and that others pose real problems of effectiveness.

Furthermore, the technique may help to reach a compromise or define a coalition of views, but it does not dictate the individual or collective judgment of the partners. Decision makers often prefer methods of this type because since they are involved in the process through a relatively simple technical framework.

Despite these factors, in the domain of evaluation in the strict sense of the term, multicriteria analysis is seldom used for purposes other than those closely resembling decision-making aid and, in particular, the ex ante evaluation of transport infrastructure projects.

However, specific problems of implementation may limit the use of multicriteria analysis, or require the presence of experts. In addition, this technique is not always used in an interactive way, as it should be, and tends to fix criteria that are, in reality, fluid.

Bibliography

Saaty T.L. (1984), Décider face la complexité, Paris: Entreprise Moderne d'Edition,

Schärlig A. (1990, 2me d.), Décider sur plusieurs critères, panorama de l'aide la décision multicritère, Lausanne: Presses polytechniques et universitaires romandes,

Roy B. et Buyssou D. (1993), Aide Multicritère la décision: Méthodes et Cas, Paris: Economica,.

12. Observation techniques

Description of the technique

Observational techniques, a form of naturalistic inquiry, allow investigation of phenomena in their naturally occurring settings. Participant observation is where the researcher joins the population or its organisation or community setting to record behaviours, interactions or events that occur. He or she engages in the activities that s/he is studying, but the first priority is the observation. Participation is a way to get close to the action and to get a feel for what things mean to the actors. As a participant, the evaluator is in a position to gain additional insights through experiencing the phenomena for themselves. Participant observation can be used as a long or short term technique. The evaluator/researcher has to stay long enough however to immerse him /herself in the local environment and culture and to earn acceptance and trust from the regular actors.

Observation consists of observing behaviour and interactions as they occur, but seen through the eyes of the researcher. There is no attempt to participate as a member of the group or setting, although usually the evaluator has to negotiate access to the setting and the terms of research activity. The intention is to 'melt into the background' so that an outsider presence has no direct effect on the phenomena under study. He or she tries to observe and understand the situation 'from the inside'.

Observational techniques share similarities with the ethnographic approach that anthropologists use in studying a culture although typically they spend a long time in the field. Aspects of the ethnographic approach are sometimes incorporated into observational methods, as for example where interest is not just in behaviours and interactions but also in features and artefacts of the physical, social and cultural setting. These are taken to embed the norms, values, procedures and rituals of the organisation and reflect the 'taken for granted' background of the setting which influences behaviours understandings, beliefs and attitudes of the different actors.

Another form of naturalistic inquiry that complements observational methods is conversation and discourse analysis. This qualitative method studies naturally occurring talk and conversation in institutional and non-institutional settings, and offers insights into systems of social meaning and the methods used for producing orderly social interaction. It can be a useful technique for evaluating the conversational interaction between public service agents and clients in service delivery settings.

The purpose of the technique

Observational techniques can be used to collect in-depth information on a few typical situations in the implementation of an intervention. The method provides detailed, rich insights into the observable output of the intervention and the influence of context, and is sensitive to the viewpoints of the key actors and the beneficiaries. Participant observation goes further in allowing experiential access to the insiders' world of meaning.

It is a particularly useful method when a study is concerned with investigating a 'process' involving several players, where an understanding of non-verbal communications is likely to be important, or where the behavioural consequences of events form a focal point of the study. Observational methods informed by wider ethnographic approaches can also provide valuable evidence about the role of institutional and organisational processes and their effects on behaviours and social meanings.

Observational techniques are also useful when one has to observe a situation about which there is little knowledge, or when it is suspected that the same situation is understood very differently, depending on whether the point or view is 'external' or 'internal'.

Circumstances in which it is applied

Observational techniques have been used to understand the functioning of policies in education, health (e.g. diagnosis, care), justice and the criminal system, scientific research, urban transport and housing. In the evaluation framework, this technique is recommended particularly for the observation of processes of interaction between the administrators and their public.

Observational techniques appear not to have been widely used in the context of the evaluation of socio-economic programmes. Yet this technique is well suited to certain Structural Fund interventions, for example those intended for the public that are difficult to observe by means of more traditional inquiry techniques (e.g. the long term unemployed, or users of illicit drugs).

If properly employed as a non-intrusive technique, observation can be used to observe the spontaneous behaviour of populations who are reluctant to accept the formalism of questionnaire surveys or provide reliable information. It is the only technique available when there are serious difficulties in gaining access to the field, for example in the case of conflict within the organisations responsible for implementation or when the beneficiaries' behaviour is partly illegal or irregular. The technique is then particularly interesting if the evaluation is launched with a view to amending the rules or regulations suspected of being ineffective.

Observational methods can be used to observe the results of an intervention of which the functioning is not well known. It is particularly useful when it is suspected that those implementing the intervention and those members of the public receiving or benefiting from it, do not perceive reality in the same way.

On the other hand, the technique is time-consuming and generates a lot of data that requires detailed processing and analysis. The use of structured observational frameworks can help to overcome this limitation, and also permit data to be aggregated and generalisations made. The technique also requires considerable skill on the part of the researcher to absorb and reflect accurately the behaviour of the key actors, and it may take time for the researcher to 'melt into the background' and therefore for participants to behave in a normal way.

The main steps involved

Observational methods generally involve the following steps.

Step 1. Choice of situations for observation

The settings for observation are defined in advance in relation to the interests of the evaluation commissioners and other key stakeholders. They consist of settings of interaction or of negotiation between public actors and the beneficiaries of the evaluated policy. The researcher negotiates access to the sites of observation with the relevant parties (informally, in the case of participant observation).

Step 2. Observation

The observer observes the course of interaction, taking care to disturb the behaviour of the actors as little as possible. This work consists of note-taking and audio-visual recordings (as discretely as possible). The observer can take notes away from research subjects or immediately after the visit.

This step cannot be limited to simple observation but must be complemented by organisational or institutional analysis so as to identify the ways in which social, cultural and physical features of the setting impinge on relations between the actors. The observer must record as much information as possible and capture an insider view of the setting.

Step 3. Analysing the material

One approach to processing the material gathered is to analyse the events observed in terms of characteristic sequences. Each recording is 'cut up' just as one would edit a film into sequences.

The observer identifies the 'evaluative assertions', that is to say, the sentences which convey an explicit or implicit value judgement. Typical sequences and their analysis are concentrated on these assertions, and reveal the way in which the policy is judged in the field. Used in this way, the tool can shed important new light on the validity and effectiveness of the policy.

Step 4. Analysis of typical sequences with the actors

The typical sequences and assertions are rewritten or modified to make them anonymous. They are then given to representatives of the people observed, for the purpose of collecting their comments and reactions. This step serves to verify that no bias has been created by taking the sequences out of their context. It gives, for each sequence, keys for interpretation which are recognised and validated by the 'community' under study.

Strengths and limitations of the technique

Observation is a generic method that involves the collection, interpretation and comparison of data. It shares these characteristics with the case study method. It is therefore particularly well suited to the analysis of the effects of an intervention that is innovative or unfamiliar, and especially the clarification of confounding factors that influence the apparent success or failure of the interventions evaluated.

Observational techniques serve to reveal the discrepancy between the way in which public interventions are understood high up at decision-making level, and the way in which it is understood in the field; it highlights the interpretation made of it by individuals in an operational situation.

The observation is generally limited to a small number of settings.

It is based on spontaneous or naturalistic data, gathered by an independent and experienced observer. The reliability of the observation depends to a large extent on the professional know-how of the observer-analyst. It is however possible to introduce a structured observational template that can be used by less experienced researchers, when gathering data across a large number of settings.

Despite its advantages, observation requires meticulous preparation to enable the observer to fit into the observed context without disturbing anyone, as well as considerable time for data collection. making it an expensive method.

The technique allows data to be gathered in difficult situations where other survey techniques cannot be used.

A major strength of using observational techniques, is that they can capture unexpected data which other methods can miss. The researcher does not define categories of data before going out into the field but is open to "what's there" - the theory emerges from the data on the ground rather than pre-defined theory influencing what data is collected.

The extent to which the observer can be present without disturbing or influencing research subjects is never nil; it is usually recommended that observers maintain self-awareness about how they impact the environment they are researching and to take account of it in their data collection. In participant observation the researcher aims to become part of a community or environment rather than maintaining a detached status.

Bibliography

Coulon, A. (1987) L'ethnomethodologie. Paris: PUF, Que sais-je?

Garfinkel, H. (1967) Studies in ethnomethodology. Englewood Cliffs, NJ. Prentice Hall. A return to the origins of ethnomethodology.

Jorgensen, D.L. (1989) Participant observation: a methodology for human studies. London: Sage Publications. Applied Social Research Methods Series, No 15.

13. Priority evaluation method

Description of the technique

The priority evaluator method (sometimes known as the priority-evaluator technique) is based on the simulation of choices in a market place and usually involves the use of social surveys to collect information. Respondents are allocated a hypothetical budget and are offered a set of items they could purchase, each with a hypothetical price. Values are then derived, according to the preferences given by respondents in spending their budget on the items.

The purpose of the technique

The priority-evaluator technique was developed as a way of involving the public in decisions about complicated planning issues. The method is an attempt to combine economic theories with survey techniques in order to value unpriced commodities, such as development or environmental conservation. It is used to identify priorities in situations where there is likely to be a conflict of interest between different people or interest groups, and the choice of any option will require a trade-off. In this respect, as an evaluation tool, the priority-evaluator technique is closely related to cost-benefit analysis and environmental economics.

The priority evaluator technique is also used to collect information on stated preferences, usually as part of the work to design initiatives which will best meet the aspirations of the intended target groups (for example, to assess travel preferences amongst an identified market of travellers, or to identify preferences for childcare amongst parents, as in the example below). In this respect, the method can be used to tailor decision making to the value-judgements held by the intended target groups in general (turned into preferences when people exchange things in the marketplace), thus avoiding decision making based on those people with most influence in the decision making process who may have a specific set of values or attitudes.

Circumstances in which it is applied

The method was pioneered by Hoinville and Berthoud in the 1970s, and was used to value travel time, road safety, vehicle pollution and vehicle congestion in London.

The method is most commonly used in relation to environmental evaluation studies to value a non-marketable environmental good. It is used when policy makers wish to solicit public views on issues such as:

- Community reactions to changes in the characteristics of the environment;
- The relative size of benefit of using the environment in a certain way, compared to non-use;
- The best quantity of environmental effects.

For example, O'Hanlon and Sinden (1978) used the technique to value existence value (defined as the benefit derived from the knowledge of the presence of a number of species), naturalness of the environment and option value (defined as the probability of seeing given species) in new South Wales, Australia.

The National Centre for Social Research (NCSR) in Britain used the method in the 1970s to measure residents' trade-offs between competing planning goals for their locality.

The main steps involved

A broad distinction can be drawn between the use of the method to identify preferences and use to measure behavioural responses. The priority evaluator technique mainly applies to the former scenario, and is designed around the identification of a set of options comprising varying levels of a given set of attributes. The basis of the technique is to let the respondent devise an optimum package, given a set of constraints. The method allows the research to identify the cost of moving from one level of each attribute to another, and the respondent is invited to choose the best package, given a fixed budget to spend. The analysis is based on neo-classical microeconomic assumptions about consumer behaviour (eg. the equation of marginal utility for all goods), thus arriving at respondents' ideally balanced preferences, constrained financially, but not limited by the imperfections and limitations of the market place. For example, respondents could be offered five 'goods' in three different quantities (say 1,2 and 3 units). Therefore, they are provided with fifteen possible choices, each reflecting a relative price for each good. The experiment is repeated for each individual with different relative prices until the ratio of the observed frequency of selection, to the expected frequency, is one, for all fifteen combinations (thus revealing the true preferences and marginal valuations).

In research focusing on environmental questions, respondents are sometimes offered various combinations of options of conventional goods and environmental amenity at assumed prices.

Strengths and limitations of the technique

The main strength of the method is in providing a scientific basis and a comparable scale to evaluate aspects of the current situation against an ideal scenario, and to assess preferences. Obviously, this is important since without such a method assessment of social, developmental or environmental conditions becomes highly subjective.

The main difficulties are in the application of the method. There is a need to identify values to several sets of choices to obtain the values, and this can be problematic. For instance in the example used on childcare, some parents may wish to reduce the amount of childcare received and so improvements would save points in this situation. The analysis also relies on statistical analysis of the results and this makes it less attractive than other techniques (for example, Stated Preferences technique).

Because it is based on the use of surveys of respondents, the method is also relatively costly and subject to sample bias issues, and design bias. In particular, where the respondent is prejudiced by the proposed approximate value allocated, or where there is inadequate detail on the effects discussed and or misleading statements. There is also potential for hypothetical bias where the decision posed in the question does not involve real market behaviour and there is no real incentive to think about and give answers that reflect their valuation. In some instances, for example in relation to environmental amenity, respondents may have imperfect information and a lack of experience of the impact on the utility being offered to them.

Bibliography

Hoinville, G. & Berthoud, R. (1970) Identifying Preference Values: Report on Development Work, Social and Community Planning Research, London.

Hufschmidt, M.M., James, D.E., Meister, A.D., Bower, B.T., & Dixon, J.A. (1983) Environment, Natural Systems and Development-An Economic Valuation Guide, Johns Hopkins University Press, Baltimore.

O'Hanlon, Paul W. & Sinden, J.A. (1978) 'Scope for valuation of environmental goods comment', Land Economics, vol.4, no.3, pp.381-387.

Hinds and Park, National Centre for Social Research (2001), Parents' Demand for Childcare in Scotland: Report for the Scottish Executive

14. Regression Analysis

The purpose of this section is to make regression analysis intuitively accessible to evaluators with a minimal knowledge of statistics. Thus, the text is focused on the interpretation of the regression estimates rather than on the technical steps needed to obtain them. Moreover, since we are ultimately interested in using regression for the purpose of evaluating the impact of policies, we are particularly concerned about drawing a plausible *causal* interpretation of the regression coefficients, which is not always possible.

In its essence, a regression is a way to summarize the direct relationship between an outcome and a set of explanatory variables, calculating for *each* explanatory variable its "net influence" in explaining and predicting the outcome—that is, net of the influence of all the other explanatory variables. We are mostly interested in the special case in which one of these explanatory variables *represents the policy we want to evaluate*. Then we are mostly interested in the influence of ***the policy variable*** on the outcome, while the other variables take backstage. Nevertheless, omitting them from the regression might distort the estimates of the impact of the policy variable.

The typical textbook rendering of regression begins with a graphical representation of the relationship between two continuous variables—that is, taking many possible values—such as food consumption and income, or income and years of education, or GDP per head and Structural Funds expenditure per head. The data can be represented by a cloud of dots drawn in a two-dimensional space. The (linear) relationship between the two variables is represented by a line fitted through the dots, called the *regression line*. An example is in Figure 1, looking at R&D expenditure and firm size, measured by the number of employees. How should one interpret the slope of such regression line? As far as the regression has a descriptive purpose, the slope represents the change we observe on average in the dependent variable (R&D expenditure) for a unit change in the explanatory variable (in this case, one additional employee). But there is a deeper cognitive ambition lurking behind the slope of a regression line. The slope coefficient is often interpreted as demonstrating the causal link between the two variables, by calling it "effect". For example, Rabe-Hesketh and Skrondal (2008) write "we use the term effect casually (not necessarily causally) as a synonym of coefficient" (page 12). Since we deal with the impact of public policies, we cannot afford to be too "casual". We'd rather be "causal", but only when the information we have warrants such a conclusion.

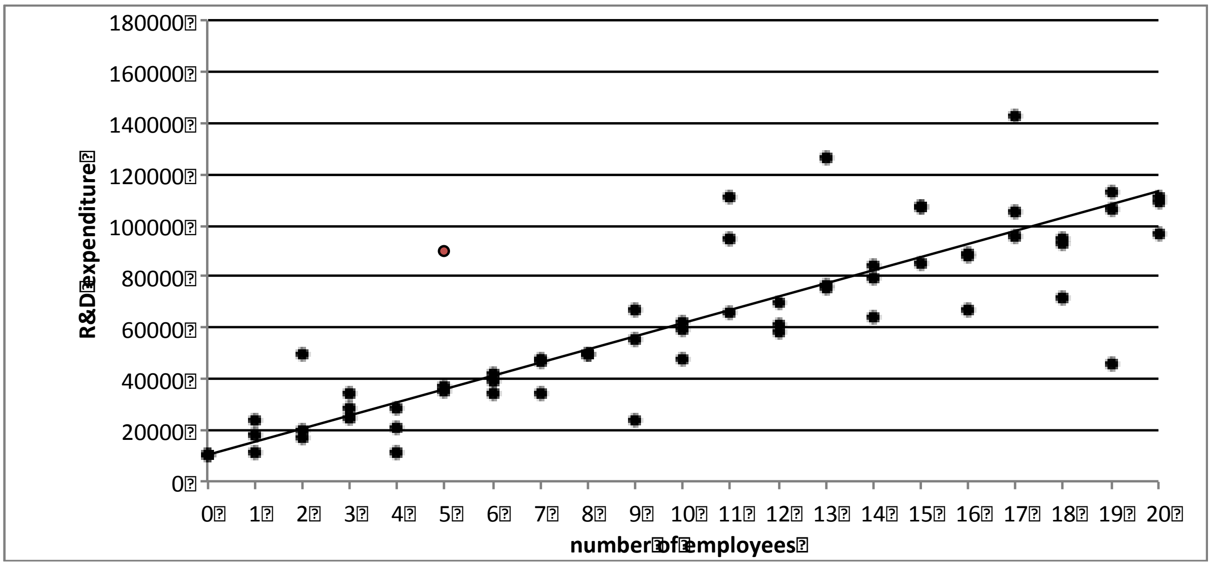
R&D Expenditure and Firm Size

Our first use of regression analysis is *descriptive*—to describe the relationship between R&D expenditure and firm size, represented by the number of employees. We are not yet concerned with establishing any causal relationship, we want to make a simple *prediction*: "What level of R&D expenditure do we *expect* from a firm with N employees?" This measure can be useful, for example, for a tax authority in a country in which enterprises can claim a tax credit for R&D expenditure, but this expenditure is not reported in the balance sheet. Thus, the tax authority can run a special survey to collect data on the relationship between size of the firm and R&D expenditure, and then make a prediction for any firm, given that the number of employees is normally observable in payroll tax databases.

The data are represented in Figure 1: on the vertical axis we put R&D expenditure (the dependent variable), on the horizontal axis the number of employees (the independent or explanatory variable). The data come from a imaginary sample of very small firms, in which the number of employees vary from 1 to 20, and the R&D expenditure range from 5,000 to 140,000 euro.

This section is based on guidance developed for DG REGIO by Professor Alberto Mostrini, Prova

Figure 1: R&D expenditure and firm size in a sample of small firms



Each dot represents a firm. For each value of the number of employees, there is some variation in the observed values of R&D expenditure. For example, among firms with 5 employees, we find values ranging from 40,000 to 80,000 euro and among 19-employee firms, values range from 40,000 to 140,000. This variation is caused by all the other determinants of R&D expenditure besides firm size, some of which are observable (and can be added to the model: see multiple regression few pages down) while other will remain unobservable.

The relationship between the two variables in Fig. 1 is clearly positive, not unexpectedly. The straight line that best approximates a *linear* relationship, called *regression line*, is fitted through the dots on the basis of a mathematical procedure known as “ordinary least squares (OLS)” (technically, it minimizes the sum of the squares of the vertical segments between the fitted line and the actual values of the dependent variable: we will return to this point in Section 5). Once estimated by OLS, the regression line in Figure 1 is represented by the following equation:

(1) Expected R&D expenditure = 10,000 + 5,000*number of employees

This means that *firms with zero employees* are expected to spend 10.000 euro in R&D (the intercept of the regression line), and that ***each additional employee*** is associated normally with an additional 5,000 euro in R&D expenditure (the slope of the regression line). We can *make predictions* using these values. We can predict that firms with 17 employees are expected to spend – that is, they will spend on average – 95.000 euro for R&D. What is the use of such predictions? As in the example given above, it might be useful to know how much small firms presumably invest in R&D, but this information is missing for many firms. After we estimate the regression line on a sample of firms for which we observe both variables, we can predict the average R&D expenditure also for those firms for which such value is missing, as long as we know the number of employees.

Note that we cannot—on the basis of the data we have available—plausibly claim that there is a *causal* relationship between the two variables: we do not know whether the number of employees is stimulating the level of R&D expenditure; or whether it is the propensity toward R&D activities that motivates the growth in the work force; or whether there is a third variable, for example the ICT nature of the firm that drives both growth in the work force and R&D expenditure. We can only say that there is an *association* between the two variables, that they are positively correlated. As the old adage goes, *correlation is not causation*.

Since we have not identified a causal relationship, *we cannot predict the impact* of increasing the number of employees. We cannot say "it is possible to increase the R&D expenditure of firms by inducing them to hire more employees". It would likely be a disappointment trying to implement such policy.[14]

The Impact of R&D Subsidies on R&D Expenditure

Let us say that now we do want to *measure the impact* – in the sense of causal effect – of giving subsidies to small and medium size enterprises (SMEs) with the goal of increasing their R&D expenditure. Under this programme, firms are eligible for grants that cover a fraction of the costs of an R&D project. The question we are ultimately interested in answering is "what is the effect of receiving the grant on R&D expenditure?", which is more challenging than simply asking "by how much do we expect R&D expenditure to differ on average between firms that receive a grant and those who do not?"

We simplify a bit the matter by representing the programmme as a *binary* indicator (receiving or not receiving the grant), which can be observed for all *eligible* firms. This simplification is less bothersome in case the size of the grant is fixed--say, 16,000 euro.   The outcome of interest—that is, the variable on which we want to measure the impact—is again R&D expenditure.

Let us suppose that the data show a *difference* of 25,000 euro between the average R&D expenditure of eligible firms who applied for a grant (100,000 euro) and of those that did not (75,000 euro). Can we conclude on the basis of these data *alone* that the grant *causes* an increase in expenditure of 25, 000 euro? ***Obviously, we cannot***. For sure, 25,000 euro is the average difference *associated* with the take-up of the grant. However, we cannot claim that the *observed* difference in R&D expenditure is *due to* the grant. The reason is that firms *applying* for the grant are likely to have higher R&D expenditure *even in the absence* of the grant. *This is a crucial point.*

Because of the *sorting process* by which firms self-select themselves by applying or not for the grant, many differences plausibly exist between recipients and non recipients, and these differences would plausibly generate differences in the R&D expenditure even if the grants were never actually paid out. These differences in pre-treatment characteristics create the so-called "selection bias problem".

[14] This argumentis known in economic policy as Goodhart's law ("Any observed statistical regularity will tend to collapse once pressure is placed upon it for control [ie policy] purposes" and is related to the famous Lucas critique ("a change in policy will systematically alter the structure of an econometric relationship")

Selection bias

This situation typically arises when participation in a programme is voluntary—as in our example, applying is a choice made freely by the firms *eligible* for the grant. If firms that invest more in R&D are also more likely than others to apply for the grant, the true effect of the grant is likely to be much *smaller* than the observed 25,000 euro. At the extreme, the true effect could be zero (or even negative!), and the observed 25,000 euro difference be due entirely to the *sorting process*.

A general principle applies to any attempt to estimate the impact of policies and programmes, whether by regression or by similar statistical tools: the ***observed*** difference in outcomes between treated and non treated units can be viewed as *the sum* of the *true causal effect* and the pre-treatment difference generated by the selection process that sorts units into treated and treated.

observed differences = true effect of the treatment + differences due to selection

We do observe only what is on the left hand side. The decomposition on the right hand side is not observable directly, but it is the implication of a plausible logical argument. The regression does *not* separate the two components, but it might help reducing the *differences due to selection.*

The only situation in which the observed difference in outcomes (most likely) coincides with the true causal effect occurs when the treatment is *allocated randomly*: in our example, this would require assigning the grant to eligible firms through some kind of *lottery*. The random assignment would guarantee that there are no systematic pre-treatment differences between the two groups. Thus, if the outcome -- observed after the experiment has been implemented -- differs on average between beneficiaries and non beneficiaries, we can conclude that the policy has an average *effect* equal to the observed difference.

Why would one allocate grants randomly? Simply to get a robust and credible estimate of programme effects, with the purpose of replicating the policy in the future with the some expectation of obtaining similar results. However, the use of randomization is fairly limited, particularly with enterprise support policies. In the absence of randomization, we must use one of the existing non-experimental methods, to disentangle causal effects from selection bias. Regression is a typical non-experimental method (though there are arguably better ones – see for example the sections on propensity score matching and other such techniques)

Regression

The simplest possible form of a regression equation is the following:

$$Y_i = \alpha + \beta T_i + u_i \quad (2)$$

where:

Y_i represents the dependent variable – the R&D expenditure for each *i-th* firm in the three years following the receipt of the grant;

T_i represents the treatment variable, which takes the value of 1 if the firm receives the grant, zero otherwise.

α is the “intercept”, representing the average value of Y when no grant is received (T=0);

β is the "slope", which coincides with the difference between the average R&D expenditure for grant recipients and non-recipients.

u_i is the so called error term and represents the influence on the dependent variable of all the unobservable factors.

Note that Y, T and u are indexed by the subscript "i", because they vary from firm to firm, while α and β are constants. The most problematic element in the equation is the term $u_{i,}$ which represents *all the other factors* influencing the Y for each i^{th} firm: these factors are – by definition – *unobservable*. It turns out that what *we assume to know* about u determines largely how the regression results can be interpreted. An innocuous assumption is that the mean of u is zero. By contrast, the second assumption at the heart of regression model is often implausible, and is made out of convenience: it is the assumption that T and u are *independent*, or that the average of u does not vary with the value of T (mean-independence).

This assumption is true only if it is plausible that the selection process determining participation in the programme, represented by T, is not related to the outcome variable Y. In our case, this is equivalent to assuming that the propensity to apply for a R&D subsidy is not related to the propensity to invest in R&D. In general, this is likely not to be true. *However, this condition holds true when the grant is assigned randomly*. In all other cases, it represents a non-testable assumption (and often, as in our case, an implausible one). However, this assumption is routinely made in most applied work.

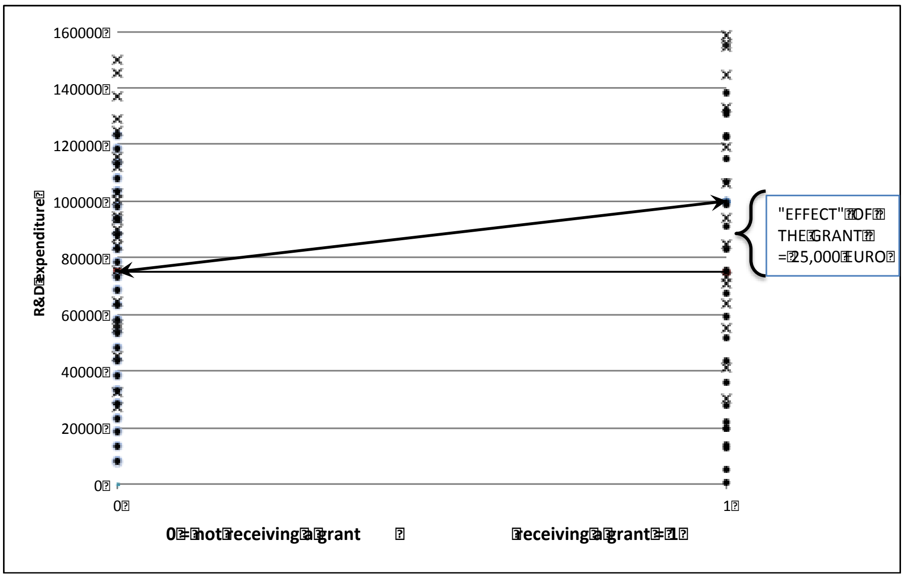
Once equation (2) is estimated with the data (we do not discuss here the issues involved in estimation), we obtain the following values for the two coefficients:

(3) Expected R&D expenditure = 75,000 + 25,000*T

The R&D expenditure is expected to be 75,000 when T=0, and 100,000 when T=1. The difference between the two values is—as the one we obtained comparing means—25,000 euro. However, the fact that the slope of the regression line is equal to 25,000 euro ***does not imply*** that the receipt of the grant ***causes*** a 25,000 euro increase in R&D expenditure.

If earlier we had reasons to ***doubt the causal interpretation*** of the *difference in means* of 25,000 euro, by the same token we should doubt the causal interpretation of the regression estimates, since the two estimates are based on the same information, and are indeed numerically identical. The vertical axis in Figure 2 represents R&D expenditure, while on the horizontal axis there are only the two values of the binary grant variable. The observations are "piled up" on the two values of T. The conditional averages are indicated by the arrow heads. The vertical distance *between* the two marks corresponds to the difference of 25,000 euro. The line bounded by the two arrow heads is the ***regression line*** that represents equation (3). The *slope* of this line has also a value of 25,000 (it is important to realize that the horizontal distance between the two "piles" is equal to 1).

Figure 2: Regressing R&D expenditure on grant recipient status



Even though there are two types of firms, indicated respectively with an X or a dot, we did not make use of this information. If we want to distinguish the two types of firms, we must add a variable to the equation. In turns, this requires the use of *multiple regression*.

4. MULTIPLE REGRESSION

Let us consider the possibility that the firms differ substantially in their R&D strategies according to whether they operate in the ICT sector or not. More precisely, ICT firms have *both* a higher probability of obtaining a grant and a higher level of R&D expenditure than non-ICT firms. The first situation is evident from Table 1, that shows ICT firms having a higher propensity to apply for and receive a grant (75%) than non-ICT firms (25%). Such higher propensity of one group over the other is not enough to create the problem of selection bias. It must also be true that being an ICT firm is associated with higher R&D expenditure. The second situation is evident from the last column of Table 1. ICT firms spend on average 40,000 euro more than not-ICT firms.

Table 1. Grant recipient status and R&D expenditure by the ICT status of the firm

	Fraction applying for a grant	Average R&D expenditure before the subsidy
Non ICT firms	25%	70,000
ICT firms	75%	110,000

In this set up, the ICT status, if omitted from the analysis, becomes a "confound". Given two variables, A and B, a confound is a third variable C causing both A and B: if omitted from the model, such omission *confounds* the interpretation of the relationship between A and B. The causality might go in the opposite direction: firms do not have higher expenditure because they receive a grant, but they apply for the grant at a much higher rate if they belong to the ICT sector, whose firms typically spend more on R&D.

This situation takes on difference names according to the discipline and the type of information that is missing: confounding, spurious correlation, omitted variable bias, endogeneity, selection bias. In the domain of policy evaluation, the most common label for this problem is *selection bias*.

The positive relationship we observe between grant status and R&D expenditure (fig. 2) could be due entirely to *selection bias* and not to the ability of the grant to influence firms'behavior. *When analyzing one relationship at a time,* we cannot help attributing the observed difference in Y entirely to either one of the explanatory variables. However, what we would like to do is to calculate the effect of each variable "net" of the effect of the other: in other words, we would like to "hold constant" one variable while letting the other one take on different values, and record the "effect" of the other one. Then we would like to record the other "effect" on Y. *The operation of holding constant one variable, while estimating the effect of the other, is what multiple regression is essentially about.*

The following regression equation:

(4) $$Y_i = \alpha + \beta T_i + \gamma \, ICT_i + u_i$$

represents the simplest set up for a *multiple regression,* because it contains only two binary explanatory variables. As before, Y_i is the dependent variable and α is the intercept, while β is the "slope" or "effect" of receiving the grant, and γ = the "slope" or "effect" of being an ICT firm. Finally, u_i= the term representing all other factors influencing Y for the i^{th} firm, besides grant and sector.

ICT is defined as a "control" variable, to underline the fact that we are not interested in its effect *per se*, but we introduce it to prevent the distortion that would be caused by its omission. The assumption needed in this case is that, conditionally on ICT, the mean of u does not vary with T.

Thus, we assume that, "controlling" for ICT the selection bias problem has been solved, and that the remaining unobservables do not vary systematically with the treatment. This remains often an implausible assumption, even with a long list of "control" variables included in the regression. What is crucial is that ***all*** the determinants of the selection process are included among the control variables. This goes under the name of Conditional Mean Independence Assumption. Only when this condition is satisfied, we can claim that the regression estimate of β represents the average causal effect of the policy.

In our example, the estimated equation is:

(5) Expected R&D expenditure = 75,000 + 10.000*T_i + 15.000*ICT_i

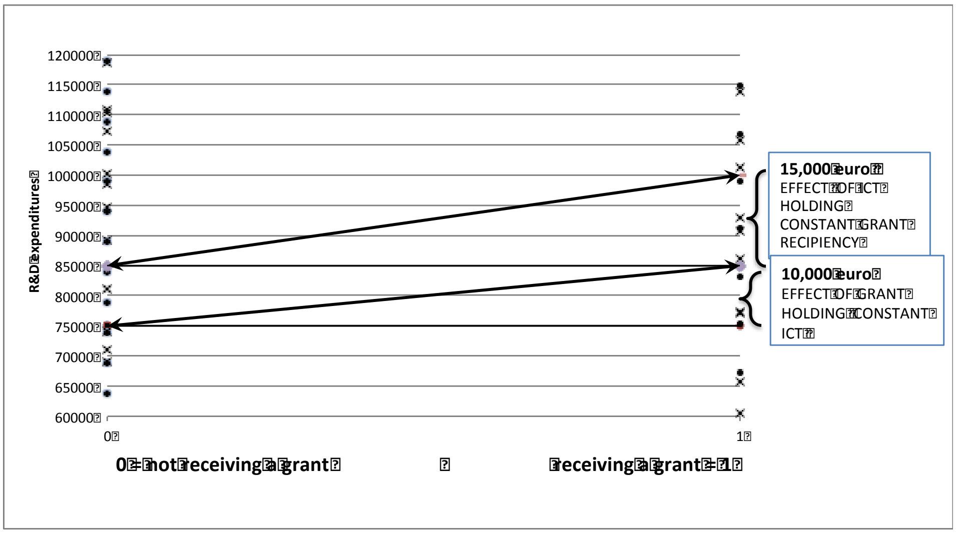
Now the "effect" of the grant is reduced to 10,000 euro, down from 25,000 in the simple regression. On the other hand, the "effect" of being an ICT firm, holding constant grant status, is 15,000 euro. Figure 3 provides a graphical illustration of multiple regression in the case of two binary explanatory variables. We no longer have a cloud of dots, the Y values "pile up" above the zero and one values of the x-axis variable, in this case grant status. Despite this unusual set up, we can still draw a regression line, with the following interpretation. The effect of the subsidy is obtained by "holding constant" ICT, that is, by separating ICT firms (indicated by the small x's) from the non-ICT firms (indicated by the dots). Such separation creates four "conditional averages". The conditional average is a formal way to express the expected R&D expenditure using the symbol E for expected and the symbol | for "given the value)

- E(R&D expenditure |T=0, ICT=0) = 75,000
- E(R&D expenditure |T=1, ICT=0) = 85,000
- E(R&D expenditure |T=0, ICT=1) = 90,000
- E(R&D expenditure |T=1, ICT=1) = 100,00

The difference between the averages on Y conditional on ICT = 1 is 100,000 – 90,000 = 10,000 euro. The difference between the averages on Y conditional on ICT = 0 is 85,000 – 75,000 = 10,000 euro (The identity between these two values in somehow forced on the data by the omission of an interaction term. See Section 4). Thus, when ICT is held constant, we are able to isolate the net effect of the subsidy. Graphically, this is represented by the lower of the two slopes in Fig. 3. The lower regression line goes through the points with coordinates (0, 75,000) and (1, 85,000).

Analogously, the difference between the averages of Y conditional on T=0 (or on T=1) represents the effect of ICT holding constant the subsidy, and is equal to 15.000 euro. Graphically, it is represented by the higher of the two slopes. It should be stressed that the simplicity of these manipulations is due to the fact that there are only *two binary* explanatory variables. Things get more complicated as soon as we add a continuous variable.

Figure 3: Regressing R&D expenditure on grant recipient status and ICT status



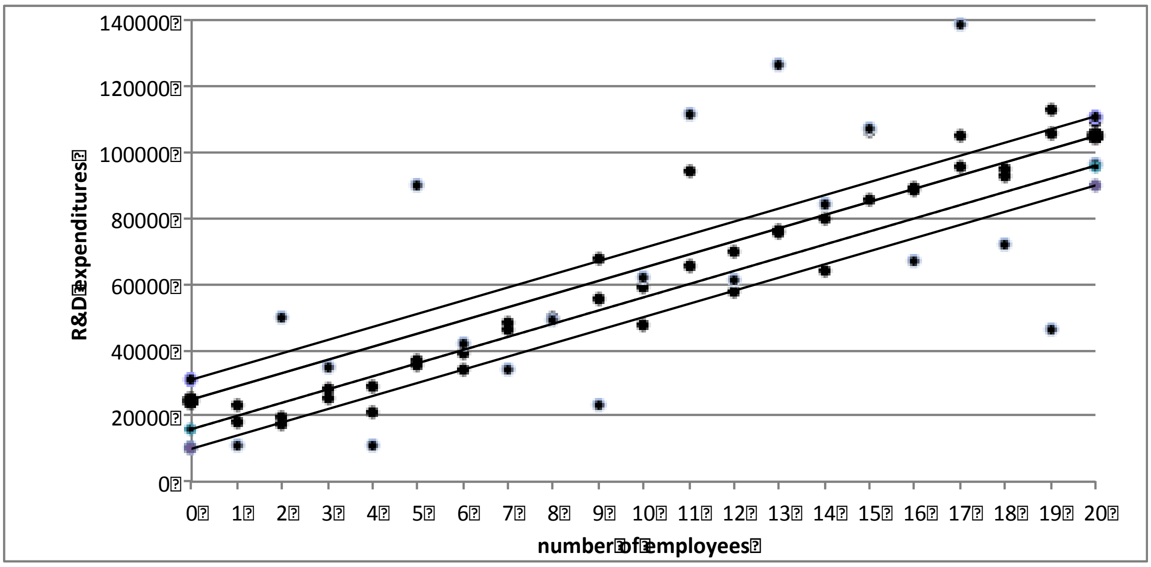
Adding a continuous control variable

Now we add to the list of “controls” a continuous variable, the *number of employees* in the firm. It is possible that the size of the firm affects both the likelihood that the firm receives the subsidy as well as its R&D expenditure. If this were the case, omitting firm size from the equation would bias upward the effect of the subsidy. We obtain the following estimates of the coefficients:

(6) E(R&D expenditure|T, ICT, SIZE) = 10,000 + 6,000*T_i + 15,000*ICT_i + 4,000*$Employees_i$

The following graph represents equation (10). The expected R&D expenditure is plotted against the number of employees, allowing the intercept to shift according to the four values for T and ICT. There are four distinct intercepts. The common slope of the four lines still represents the effect of an additional employee on R&D expenditure.

Figure 4: Regressing R&D expenditure on grant recipient status, ICT status and firm size



The following are the values of the four intercepts, they can be easily derived from equation (6).

T=0 ICT=0	10,000
T=1 ICT=0	16,000
T=0 ICT=1	25,000
T=1 ICT=1	31,000

The parallelism between the four lines in Fig. 4 is a direct consequence of how we specify the model. Namely, since we used a purely additive specification, the effect of one variable is "added on top" of the effects of the others. To avoid this problem, we need to enrich the model by adding ***interactions*** between the variables. It is beyond the scope of this introductory material to discuss fully the use of interactions (as well as the use on non-linear terms). However, we provide some intuition by showing an interacted version of equation (6). We want to know *whether the impact of the grant varies with the size of the firm*. Note, we already know from equation (6) that size and grant have a positive effect. We want to know whether size modifies the effect of the grant. To this end, we add an "*interaction*" (that is, the product) between size, sector and the receipt of the grant.

Adding interactions

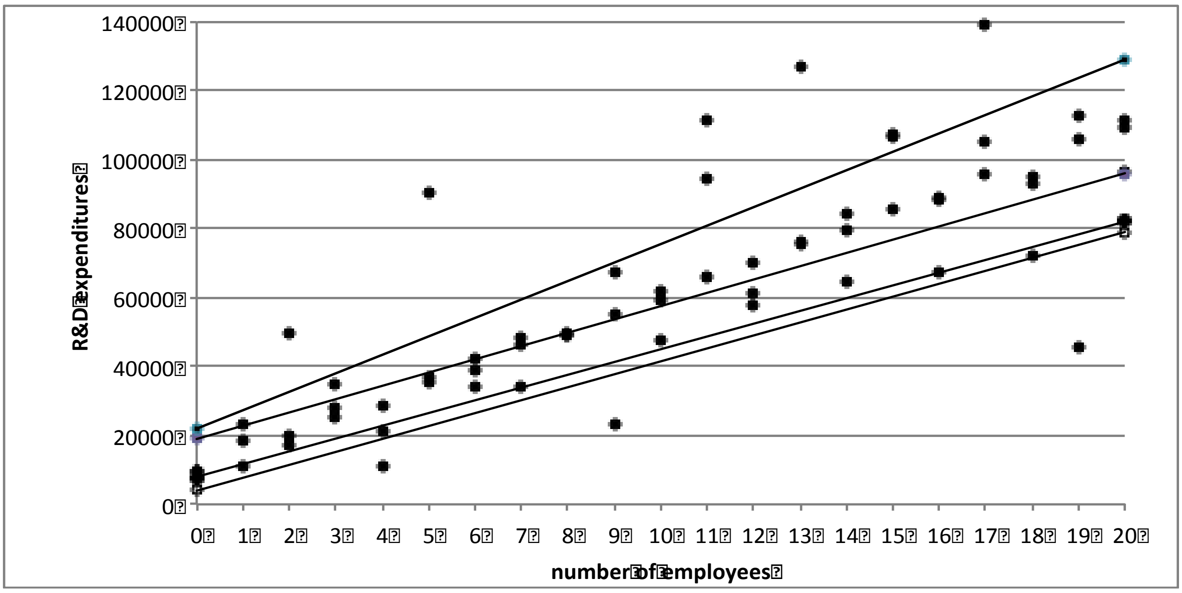
We obtain the following result:

(7) $E(R\&D \mid T, ICT, Size) = 1{,}000 + 3{,}000{*}T_i + 15{,}000{*}ICT_i + 3{,}000{*}Empl_i + 3{,}500{*}Empl_i{*}\ T_i{*}ICT_i$

The following graph represents (approximately[15]) equation (7).

[15] Actually, the model should include two other interaction terms: between T and firm size and between ICT and firm size. The representation in figure 5 corresponds to a coefficient of zero for both these additional interactions, so that the only differential effect is limited to ICT firms receiving a subsidy.

Figure 5: Regressing R&D expenditure on grant recipient status, ICT status, firm size and the interaction of grant status, ICT status and firm size



The two regression lines standing above the others are those for ICT firms. ***The two regression lines no longer have the same slope***: for each additional employee in a ICT firm, we expect the R&D expenditure to increase more rapidly for recipients than for non recipients. For the T=1 firms, the increase is 3,500 + 3,000 = 6,500 per additional employee, for the T=0 firms the increase is only 3,500. It turns out that the same pattern does not hold true for non-ICT firms: their regression lines are parallel and the effect of receiving a grant is always 3,000 euro. Interactions are the simplest way of capturing *impact heterogeneity.*

Impact heterogeneity bears important ***implications for public policy***: this is a way to provide the type of information policy-makers need for better targeting scarce resources to those who can mostly benefit for the subsidy. In this (invented) example , the policy implications of (7) would be clear: *target R&D subsidies to larger ICT firms*. If indeed the grant were given in fixed amounts of 16,000 euro, there would be a positive net gain of providing grants to all ICT firms with more than 5 employees =(16,000-1,000)/3,000. In this example, it would never pay to provide non-ICT firms with R&D subsidies.

The policy implications produced by the analysis are credible only if the model is a good one: but how do we know this? There is not an easy answer, particularly not one that can be explored in depth in this introductory text. A very partial answer is provided in the following section.

Strengths and limitations of the technique

There are two very basic, and different, questions concerning the quality of the regression estimates.

Goodness of fit – interpreting "R²"

The first issue has to do with how well the model as a whole "fits the data" or, put differently, to what extent the model explains the variability of the dependent variable. This is represented by a measure called R^2, which varies conveniently between zero and one.

A value of $R^2 = 0$ means that the regression does not explain any of the variability of the dependent variable, while a value $R^2 = 1$ means that the regression explain all such variability (if the $R^2 = 1$, all data points lie on the regression line). We will provide some intuition comparing the following two graphs.

Figure 6a: R&D expenditure and firm size with R^2 close to 1

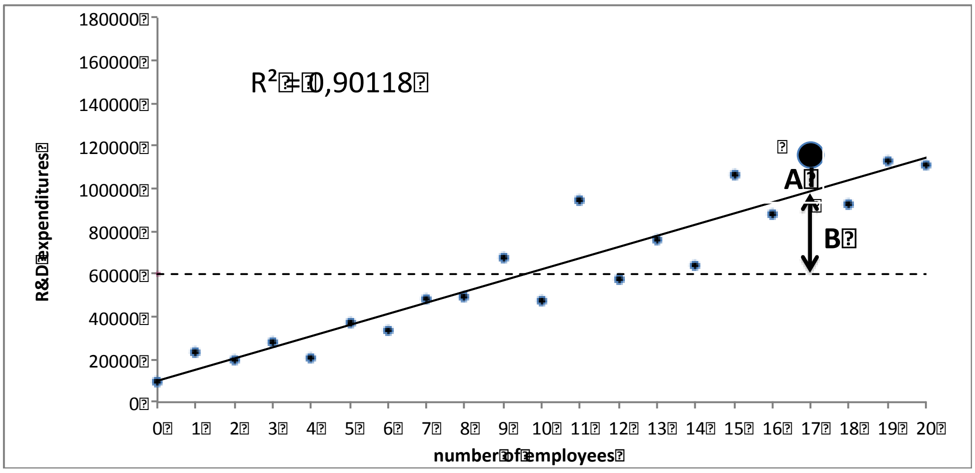
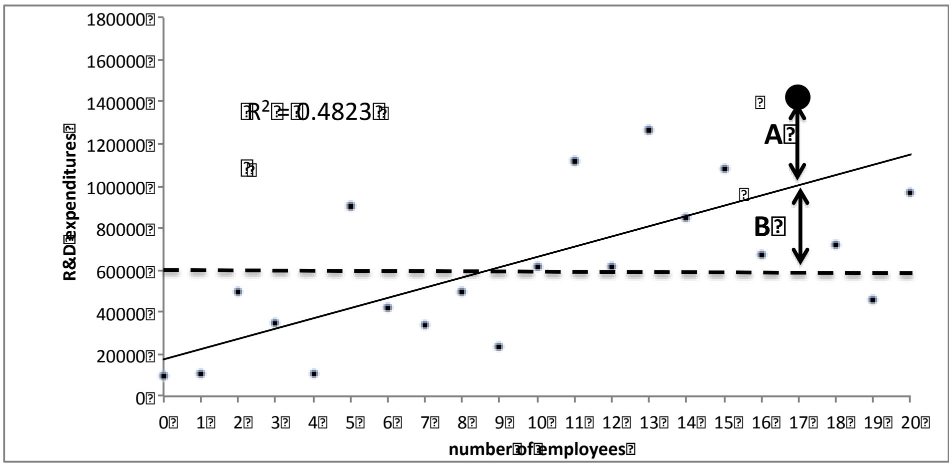


Figure 6b: R&D expenditure and firm size with R^2 much smaller than 1



The regression line is in the same position in both graphs, so that intercepts and slopes will have the same values across the two models. Figure 6a represents the situation in which R^2 is indeed close to 1: the data points lie very close to the regression line. By contrast, Figure 6b shows much more dispersion around the regression line, and the value of R^2 is less than 0.5.

How is the value of R^2 calculated? The intuition is the following. Take the observation represented by the large dot in each graph. The vertical distance between the dot and the dashed line, representing the average of the dependent variable, can be decomposed into two segments. Segment A represents the "error of prediction", also known as residual, and reflect the partially unpredictable nature of any phenomenon. The regression line has been drawn in a way as to minimize the sum of the (squared) residuals.

Segment B indicates by how much the model predicts the observation will be above the average of the dependent variable.

Segment A is very short in figure 6a, compared to segment B, while segments A and B are about the same length in Figure 6b. The value of R^2 depends on the relative size of all A and B.

More precisely:

$$R^2 = \text{sum of all } B^2/\text{sum of all } (A+B)^2$$

if we divide both numerator and denominator by N, we obtain

$$R^2 = \text{variance of Y explained by the regression /total variance of Y}$$

R^2 close to 1 means that the regression model—that is, the explanatory variables—is able to explain why Y varies across cases. When there is more dispersion around the regression line, the model does not explain much variability.

Does this imply that in the case of low R^2, the results are *not* policy relevant? We strongly emphasize that such claim is misleading. We would like to offer a strong counterargument. The degree to which the variability of the dependent variable is explained by the model has ***little relevance for evaluating the impact of the policy***. The purpose of impact evaluation is not to explain the whole phenomenon, rather it is to quantify the "net" effect of the exposure to a policy. To be sure, we would rather base our policy advice on a better fit to the data than not. However, that is not the purpose of impact evaluation.

Standard errors

The purpose of impact evaluation is to estimate with precision the slope of the variable representing the policy, making sure this variable is truly independent of the error term. "With precision" means with a standard error which is small compared to the value of the coefficient. The standard error represents the variability of an estimate (like a regression coefficient, or the difference between two means) one would obtain by drawing repeated samples of size N from the same infinite population. Standard errors play a central role in statistics, because they quantify uncertainty. The key idea is to compare the statistic obtained from the sample with its standard error. A very common rule of thumb is that the (absolute value) of the statistic should be ***more than twice its standard error***. If this condition is satisfied, we conclude that the effect of the policy is *statistically significant* (that is, is not due to chance).

Three caveats are in order. *First*, the fact of being statistically significant does not prove at all that we identified a relationship that is interpretable in a casual sense. "Causality is in the mind" as poignantly writes Jim Heckman (Heckman, 2008), "and it is not easily impressed by low standard errors", we would like to add. We can easily find examples in which a perfectly silly model has been estimated to produce statistically significant but perfectly silly results.[16]

[16] If there were a competition for Silliest Regression Model, there would many contestants, but one of the prizes would go to the following model, a regression of average change in GDP per capita on average change in EU structural funds per capita, using 11 observations. Apart from drawing conclusion based on a dozen observations, given that structural funds allocation are determined on the basis of GDP per capita, the relationship can hardly be intepretable as causal. But the authors are pleased by the results, boasting a R^2 of 0.53 and a t-statistic of 3.

Second caveat. Statistical significance is largely driven by sample size. As N grows to infinity, all differences become significant, all null hypotheses are rejected, all slope coefficients are precisely estimated. To be sure, other factors also influence the size of the standard error. The formula of the standard error of the slope coefficient can be written as follows.

$$\text{standard error of } \beta = \sqrt{\frac{var(Y)}{var(X)} \frac{1 - R^2}{N}}$$

This formula makes some intuitive sense. Given a value of the slope coefficient on X, its standard error increases with the variance of Y and decreases with the variance of X. But typically the two variances are given. Moreover, the standard error goes to zero as the R^2 gets close to 1. However, a R^2 close to 1 is a rare event. Finally, the standard error goes to zero as sample size goes to infinity. Particularly because the increased availability of large administrative databases, in the long run we will all be statistically significant.

Third caveat. We want to judge the **policy** significance of the estimates, no matter what the **statistical** significance is. The most obvious way to do so is to compare the impact with the cost of producing it. In our example, the final estimate was an average impact of 4000 euro. That is, the average increase caused by the grant among recipient firms is 4000 euro. This can be very precisely estimated or not, but it remains a very small effect, particularly if the average grant is 16,000 euro.

Concluding Remarks

The estimates of the effect of the grant on R&D expenditure vary widely in the examples we gave, from the 25,000 euro with no control variables, to 10,000 euro controlling for ICT sector, to 4,000 euro when we control also for firm size. Then we have learned that the impact of the policy is much greater for some types of firms. Have we established beyond any doubt the *true* causal effect of the grant? ***Not really***.

We only made more and more plausible (or less and less implausible) the assumption that we have eliminated selection bias. No addition of control variables or changes in the functional form of the regression equation will *insure* that one is able to overcome the selection problem. The problem is that we do not know the process that generated the data, we only observe the final outcome of that process. This (the non observability of the counterfactual) has been called the “fundamental problem of causal inference” (Holland 1986). Multiple regression is a tool to correct this problem, by incorporating all the variables that are believed to influence both outcome and selection into treatment. But regression is not the only tool in the toolbox. Evaluations often combine it with other methods (eg Difference-in-Difference) or even dispense with regression altogether in favour of methods such as Propensity Score Matching.

Bibliography

Heckman, J. “Econometric Causality”, UCD Geary Institute Discussion Paper Series (2008)

Holland, P., “Statistics and Causal Inference”, *Journal of the American Statistical Association* (1986)

Lucas, R "Econometric Policy Evaluation: A Critique", in Brunner, K.; Meltzer, A., *The Phillips Curve and Labor Markets*, Carnegie-Rochester Conference Series on Public Policy, 1, New York: American Elsevier. (1976),

Rabe-Hesketh S. and Skrondal A., *Multilevel and Longitudinal Modeling*, College Station, TX: Stata Press (2008).

15. SWOT Analysis

Description of the technique

SWOT is an acronym, standing for strengths (S) and weaknesses (W) of the organisation, and opportunities (O) and threats(T) of its environment.

SWOT is not an analytical tool per se; instead it is a way to synthesize preceding analyses and use them for developing a strategy.

In its basic form, it is a structured list of characteristics of the organisation and its environment, often used to facilitate discussions about strategic planning. It became a very widespread tool mainly because it is simple and can be applied with little preparation.

In Cohesion Policy, the method is used differently, as part of the socio economic analysis of regions. In some cases it is used as a support to identify a strategy rather than to test the "strategic fit" of a decided strategy.

The purpose of the technique

SWOT analysis was developed in the 1960s and originates from the works of business policy academics. It was designed to be used by companies in their strategic planning process.

Conceptually it relies on the concept of “strategic fit”: the idea that an organisation is successful if its internal characteristics (strengths and weaknesses) fits the external environment (opportunities and threats), and fundamental role strategic planning is ensure this fit in the long run.

Once the internal and external factors have been identified, SWOT pairs the strengths and weaknesses with the opportunities and threats, essentially giving four types of strategic possibilities:

- Using strengths to exploit opportunities;
- Using strengths to avoid or minimalize threats;
- Identify and address weaknesses that may prevent achieving objectives; and
- Identify weaknesses that make the organisation vulnerable to threats.

These possibilities are not necessarily alternatives. The key to develop a strategy is to combine these possibilities in a way that ensures the ‘strategic fit’, thus the organisation is able to maximise benefits that comes from the changing environment.

Circumstances in which it is applied

The subject of a SWOT analysis can vary as the technique is very flexible: it can be applied for a company, a product, a project, a policy or even a development programme. However, it can be used to judge the ‘strategic fit’ in relation to a certain objective, thus setting objectives precedes the SWOT analysis.

Therefore SWOT analysis should be applied to judge whether an objective is attainable or not.

The main steps involved

SWOT is a flexible tool that can be applied in various ways. Nevertheless, to maximize its contribution to the strategic planning process, some common steps are necessary:

Step 1. Identifying the objectives of the programme

As mentioned above, SWOT is interpreted in relation to objectives. This step can be interpreted as the preparation for SWOT: there is no reason to carry out SWOT analyses for unrealistic objectives.

The first step is important because the elements of SWOT can only be understood in relation to an objective: it is necessary for deciding whether a factor is beneficial or disadvantageous, and also whether it is relevant for the analysis or not.

Step 2. Identifying external factors: opportunities and threats

External factors are characteristics of future trends (that may already exist in the present) that can influence the attainment of our objectives. They generally cannot be controlled by the decision-makers, even if the programme can influence them to some extent.

The actual analysis can be part of the SWOT or outside of the exercise; the focus is on identifying them and assessing the extent to which they are favourable or adverse.

Step 3. Internal analysis of strengths and weaknesses

Internal factors are characteristics that already exist and are under the control of the decision-makers. Depending on whether they help or prevent attaining the objective, they are regarded as strength or weakness.

Similarly to the previous step, the analysis can be part of the SWOT outside it.

Step 4. Pairing external and internal factors

Four strategic possibilities are developed in this step. It is best represented in a 3x3 matrix:

	Strengths	**Weaknesses**
Opportunities	exploit opportunities	identify risks
Threats	avoid / minimalize threats	identify vulnerabilities

If the SWOT elements are very complex, they may be divided into categories (e.g. threats may be divided into environmental and economic). In this case the number pairings can be increased, e.g. pairing strengths and weaknesses against both environmental and economic threats.

Step 5. Developing the strategy

The four strategic possibilities will be combined to develop a strategy. The interactions between the different possibilities can be very complex, as they can easily complement or contradict each other.

It is important to note that this is an iterative process: during the development of the strategy, the objectives should also be reviewed. Discarding an objective because it proves to be unattainable is common, although a SWOT usually leads to modifications of the objective.

The final product of this step will be objectives and the strategy to achieve them. Depending on how detailed the analyses were, the SWOT can be used to develop different actions for more operational purposes.

Strengths and limitations of the technique

Despite its popularity, SWOT is also criticised both conceptually and in its application.

SWOT may not be feasible where the environment (or the organisation that develops the strategy) is changing very rapidly or future trends are very uncertain. In these cases SWOT can lead to “false uncertainty” and fix a strategy that may be adverse.

Carrying out a SWOT analysis has numerous pitfalls:

- The SWOT will become too long. There is a tendency to list and classify all characteristics of the organisation or the environment (even if they are not relevant for the objectives).
- Items on the list are unclear and unambiguous. People tend to use two or three words to describe a complex factor, which can lead to false interpretation if a detailed description is not included.
- There is no prioritisation of the elements: each strength, weakness, etc. that is listed has equal importance. This can be avoided by giving weights or priority to them.
- The same factor can appear in contradicting categories, and this contradiction is not analysed or put in the perspective of the objective. E.g. “low unemployment” can be considered both as a strength (low pressure on social security) and as a weakness (lack of idle labour, pressure on wages). How this phenomenon is addressed may depend on further analysis (e.g. which industries are the main source of employment), the environment (e.g. future trends in these industries), and on the objectives (e.g. if the objective is improving the environmental situation of the region or structural change in the economy).
- In many cases the SWOT contains statements that are not verified or supported by evidence but rather reflect the opinion of people working on the SWOT. This can lead to serious problems when there is a significant gap between reality and perception.
- SWOT is carried out without pre-defined objectives and it starts with step 2. The objectives are not set even at a later stage, which results in a very lengthy and difficult process when elaborating the strategy. Setting objectives outside the SWOT in a different exercise is also a

common mistake, and as a consequence, the SWOT analysis and the strategy will not be related.

- A very common problem of SWOT is that it remains a simple, although structured, list of factors – that is steps 4 and 5 are omitted. Again, the subsequent strategy (if it would be developed at all) will not relate to the analysis.

Bibliography

- Terry Hill and Roy Westbrook: SWOT Analysis: It’s Time for a Product Recall, Long Range Planning, Vol.30, No.1 (1997)
- Heinz Weihrich: The TOWS matrix - A Tool for a Situational Analysis, Long Range Planning, April, 60 (1980)