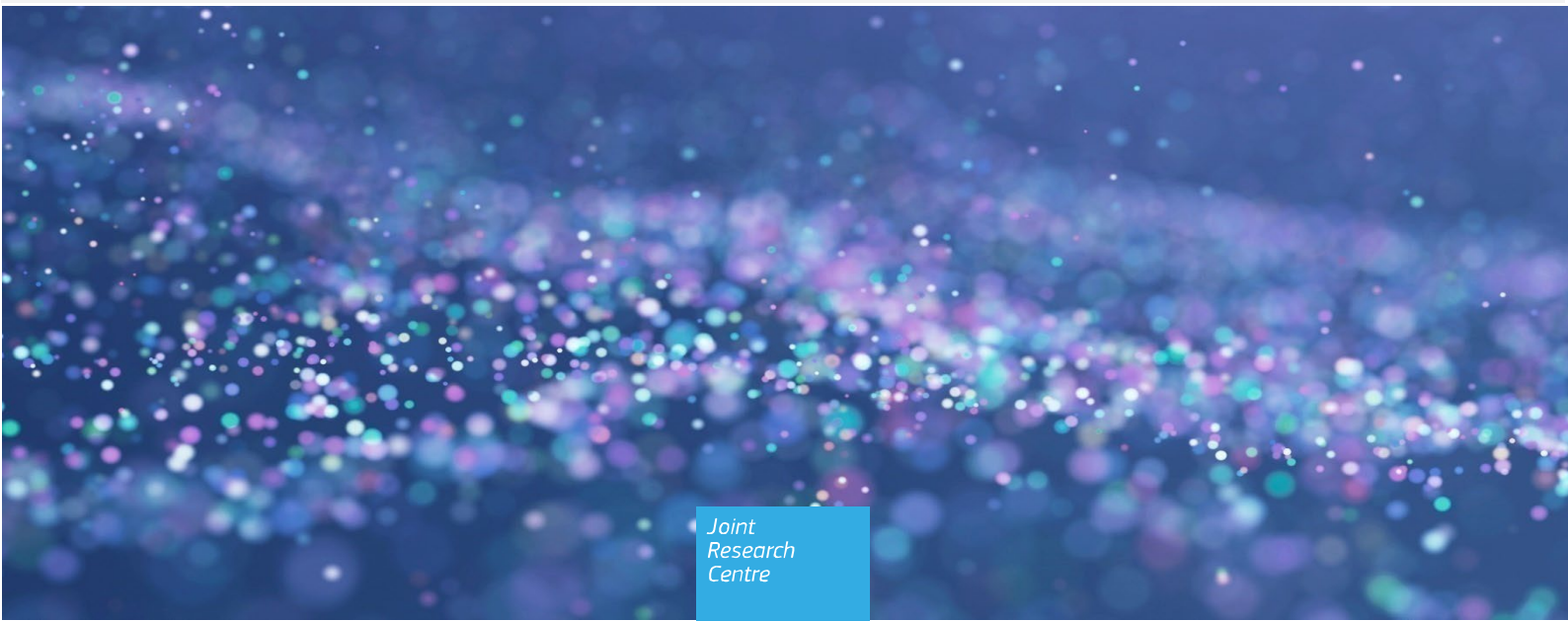
European Commission

JRC EXTERNAL STUDY REPORT

Guidebook for conducting counterfactual impact evaluations of State aid schemes for research development and innovation

Czarnitzki, D.

2023



Joint Research Centre

Contact information
Name: Prof. Dr. Dirk Czarnitzki
Address: Naamsestraat 69, box 3535 | 3000 Leuven | BELGIUM
Email: dirk.czarnitzki@kuleuven.be
Tel.: +32 16 326 906

EU Science Hub
https://joint-research-centre.ec.europa.eu/jrc

JRC135293

PDF ISBN 978-92-68-08453-3 doi:10.2760/814181 KJ-02-23-146-EN-N

Luxembourg: Publications Office of the European Union, 2023

How to cite this report: Czarnitzki, D., *Guidebook for conducting counterfactual impact evaluations of State aid schemes for research, development and innovation*, Aquaro, M., Crivellaro, E., Di Novi, C., Granato, S., Sasso, A. and Vidoni, D. editor(s), Publications Office of the European Union, Luxembourg, 2023, doi:10.2760/814181, JRC135293.

Contents

Abstract

This guidebook describes why quantitative counterfactual impact evaluations (CIE) of research, development and innovation (RDI) State aid policy schemes should be carried out, and how such CIE studies can be implemented. It discusses the motivation for governmental intervention in the market for RDI. For instance, why positive external effects are associated with RDI activities and how financial constraints for RDI arise. The guidebook then explains why impact evaluations of policy interventions are important and presents econometric methods that may be applied in CIE studies. Examples are matching methods, difference-in-difference regressions and regression discontinuity designs, among others. Some existing good practice examples of CIE studies are presented after the description of each econometric method. The guidebook also discusses data requirements for CIE studies, and suggests potential data sources that could be used in the area of RDI, such as firm-level innovation and Research & Development databases, as well as patent and trademark data. It also describes how the design of schemes may be used to identify policy effects. It concludes with suggestions on how to prepare CIEs.

1 Introduction

1.1 Economic reasons for governmental intervention in the market for R&D and innovation[1]

1.1.1 Positive external effects of R&D and Innovation activities

It is a common opinion that firms cannot appropriate all of the returns from knowledge-creating investments, such as research and development (R&D). R&D activities create information; something as intangible as information, and thus knowledge, can never be kept fully secret. Knowledge spills over to other companies that may benefit from this information. Therefore, the social returns of the initial investment are higher than the private returns in cases where others do benefit from knowledge that has been created elsewhere (i.e., the initial investor's innovation projects create positive external effects). The firm (or the inventor) will only invest in projects where the expected private return is higher than the private cost because the initial investor can only appropriate the private returns. Thus, many innovation projects may not ever be implemented, even though the social returns are high, because the private returns do not cover the private cost.

For example, if a firm has developed a new product or service that is very successful on the market, then rivals are typically able to imitate the innovation within short time periods, by business re-engineering for instance, and can sell similar products or services, thereby benefiting from the original investor's R&D investment. The increased competition will lead to lower prices and this is beneficial for consumers. The imitation could happen to such an extent that the original investor becomes unable to cover the R&D and innovation cost, and it might happen that the innovation would, therefore, not be undertaken in the first place.

1.1.2 Financial constraints

Governmental intervention is often also justified because of financial constraints for R&D investments. According to the economic theory of perfect capital markets, firms should be able to raise external capital for investments that have positive expected profits, including R&D projects. In practice, however, information asymmetries between borrowers and lenders about projects' qualities and their expected returns lead to premia that borrowers have to pay for external capital when internal capital resources are insufficient for conducting a certain investment project. The effects of such information asymmetries are most pronounced in investment projects concerning R&D as the outcome of such endeavours, something which is inherently uncertain. Thus, even if potential lenders are able to form some expectations about expected profits, the variance of such expectations may be high and, therefore, banks and other investors will require risk premia and collateral.

Providing collateral is also most problematic in cases of R&D projects because such activities mostly entail wages for R&D staff and other current costs that are immediately sunk once expensed. This stands in contrast to investments into buildings and equipment, assets that are activated in the firms' balance sheet. Tangible assets can be sold and have positive redeployment values unlike the lion's share of R&D expenses.

In 2019, according to the OECD R&D statistics, the share of labour cost and other current cost in total R&D spending amounted to 93% in Germany and Spain, 90% in Italy and 87% in France, for instance (cf. Table 1).

The economic literature on financial constraints for innovation reports that the lack of financial resources is a barrier to innovation, particularly in small and/or young high-tech firms. Such firms are relying on a continuing stream of innovative products and services in their business models on the one hand, but are also most prone to financial constraints as they typically have low internal financial resources and often low inside collateral values that could facilitate the receipt of obtaining loans on the other. Young firms also lack convincing track records on the financial market which makes it even more difficult to request external financial resources successfully.

[1] This report does not provide a comprehensive overview of R&D and innovation policies, but instead examines a selected set of the most common reasoning why we ought to publicly support R&D and innovation in the business sector as well as a selected set of the most widely applied policy schemes.

Table 1: Business R&D expenditure by type of activity in 2019 for selected EU countries (in million EUR)

	Total cost	**Capital cost**	**Current costs**		**Share of current cost**
			Labour cost for internal R&D personal	Other current cost	
France*	32,326	4,048	19,040	9,236	87%
Germany	75,830	5,595	46,524	23,712	93%
Italy	16,589	1,615	10,956	4,018	90%
Spain	8,741	575	5,285	2,881	93%

Source: OECD R&D statistics, 2019

1.2 Policies to promote innovation and the dilemma of crowding out effects

Governments of all industrialised countries (and others) find it justified, from a societal perspective, to publicly intervene into the markets for R&D and innovation, and to subsidise certain R&D projects that promise high social returns because of the external effects of R&D, described above, and the financial constraints arguments; in particular, provided that the expected social returns are much higher than the expected private returns and where the private costs are expected to be higher than the private returns.

Governments usually employ a number of different technology policies in order to incentivise innovation in the business sector. Among others, creating intellectual property rights systems such as the (i) *patent system*, (ii) *registered designs* and (iii) *trademark systems* provide incentives to invent with the benefit of promising ex-post rewards in the form of possible, temporary quasi-monopoly rents for the protected inventions or corresponding goods and services, respectively. *Patent boxes* add additional incentives through reducing corporate taxes incurred because of profits made from the sales of products or services that entail patented inventions. Exempting *R&D collaboration* among firms from anti-trust laws is also a policy that promotes innovation.[2] A policy tool that has recently received attention is represented by demand-side policies, such as *public procurement* that may or may not include contracted innovation (PPCI). The most popular, however, and the main subject of this guidebook, are policies that directly reduce the cost of R&D at the firm-level in a given year. These are *R&D tax incentives* that reduce the cost of R&D by decreasing the tax burden of R&D-performers per Euro of R&D that is expensed and *direct R&D grant programmes* that distribute public financial resources to firms that undertake selected R&D projects.

This guidebook aims to describe how to evaluate such direct R&D-cost reducing policy schemes quantitatively regarding their impacts. While the focus of this guidebook is mainly on R&D grants, given that these programs are the most commonly co-supported by EC resources in the EU member states, it also mentions particularities concerning specific challenges in evaluating R&D tax incentives.

Why evaluation?

Member States provide State aid for achieving a wide array of policy objectives in the European Union, among others, to promote research and development and innovation activities. EU State aid rules imply that aid schemes are only approved ex-ante if their expected positive effects outweigh any negative effects because policy schemes must be compatible with the common market. Examples of negative effects could include distortions of competition, inefficacy, or inefficiency. Here, ex-post evaluations become necessary In order to verify ex-post whether policy schemes actually comply with the assessment criteria (cf. European Commission, 2014).

Consider the example of R&D grant programmes. Suppose a firm has several project ideas in a given year. Which of these projects will the firm carry out? The firm will sort the project ideas according to their expected marginal returns and will conduct R&D for those projects as long as the expected marginal returns are above the marginal cost of R&D activities. The original idea of publicly supporting R&D projects is that the firm might have a project idea that does not fulfil the condition that the expected marginal returns of R&D be larger than the marginal cost, but the project's expected social returns are larger than the private returns and the private

[2] Block Exemption Regulation (EU) No 1217/2010 on the application of Article 101 (3) Treaty on the Functioning of the European Union to certain categories of R&D agreements ("R&D BER").

costs. This project would not be conducted even though it would be welfare enhancing because firms behave in a profit-maximising fashion.

If policy schemes for public support of R&D are in place, then the firm could apply with the project proposal for a subsidy to reduce its own cost of R&D, thereby making the project implementation profitable for the firm. The optimal subsidy would be the amount required to make the project just marginally profitable. If the policy scheme would distribute grants to all such projects that have high social returns, but which would not be implemented in the absence of the same policy, the policy scheme's so-called "additionality effect" would be maximal.

In practice, however, this theoretical optimum may never be reached. Once policy schemes for R&D support are in place, firms have an incentive to also apply with project ideas that they would have implemented anyway because the private costs are below the private expected returns. In the real world, it is difficult for managing authorities to identify projects that have high social returns, but also where the private costs exceed the private returns. If such projects are selected for funding, then the firms might, in the worst case scenario, just substitute public money for private money and not conduct any additional R&D project that promises further positive welfare effects for society. The economic literature refers to "full crowding-out" in such cases.

The risk of crowding-out effect may differ in practice across policy schemes and might sometimes hinge critically on many aspects of the design of the scheme (i.e., the intervention logic, the eligibility criteria, the selection criteria of project proposals, the extent and mode of public support (most R&D grant programmes are set-up as matching grants where the funding agency reimburses a certain share of the R&D project's total cost) and also the interplay with other policies that are in place in a certain country or region).

The question of whether a policy scheme is effective (and efficient) can usually only be answered on a case-by-case basis because policy schemes differ in many dimensions. Empirical policy evaluation studies have to be conducted to assess a policy scheme's success. Econometric estimation techniques for counterfactual impact evaluation studies have gained a lot of attention in policy practice in recent decades. These econometric models try to establish a causal link between participation in an R&D grant programme and observed outcomes (e.g., R&D investment, in comparison to counterfactual outcomes that would have been realised if the firm had not participated in the grant scheme) with methodological rigor. The results of such econometric estimators allows, for instance, for the rejection of the hypothesis of full crowding-out effects, and thus no efficacy, with precise statistical tests.

This guidebook, therefore, focuses on quantitative econometric counterfactual impact evaluation studies as a tool by which to assess the effectiveness of R&D and innovation policy schemes.[3]

1.3 The core evaluation of the problem

One faces an evaluation problem when investigating whether a policy scheme creates additionality, or in other words when it is not subject to (full) crowding-out effects. The core evaluation problem can be best described as a comparison of two different states surrounding the firms' participation in the policy scheme under scrutiny.

Suppose that the policy scheme supports R&D activities of firms with direct grants and that the managing authority would like to learn whether the scheme actually increases firms' investment into R&D. In this instance, one would typically like to know whether the participating firms invested more into R&D compared to a so-called counterfactual situation in which they would have not participated in the programme. In quantitative studies, one usually considers the *average* R&D spending of all participating firms, compared to what these firms would have invested on average had they not been awarded a grant from the policy scheme. The average difference in the outcome variable of interest (e.g., R&D spending) between the factual situation and the counterfactual situation of the programme participants is referred to as the *average treatment effect on the treated*.

In empirical studies, scholars can usually observe the factual situation (i.e., how much the programme participants spend on R&D in a given year). These R&D expenses should be compared to the counterfactual of no programme participation. However, the counterfactual situation cannot be observed at the same time because the firms can only be observed in the factual situation in a given year.

[3] For companion guidebooks, see Conti et al. (2021) on regional state aid and Farrell (2022) on energy state aid.

Box 1. The fundamental evaluation problem as equation

The fundamental evaluation question can be illustrated by an equation describing the average treatment effect on the treated:

$$E\left(\alpha_{TT}\right) = E\left(Y^T \mid S=1\right) - E\left(Y^C \mid S=1\right)$$

where Y^T is the outcome variable. The status S refers to the group: S=1 indicates the treatment group and S=0 the non-treated firms. Y^C is the potential outcome that would have been realised if the treatment group (S=1) had not been treated.

While $E(Y^T|S=1)$ is directly observable, it is not the case for the counterpart. $E(Y^C|S=1)$ has to be estimated.

Holland (1986) refers to this as the fundamental problem of causal inference. The estimation of $E(Y^C|S=1)$ is, thus, the main common goal of all methods for counterfactual impact evaluation studies.

2 Methods for counterfactual evaluation

Suppose we consider the example of a policy scheme for R&D grants, suppose we have data on R&D spending for the participating firms and that this exists for a sample of non-participating firms. It seems tempting to derive the average treatment effect on the treated by comparing the average R&D investment of participating firms to the R&D investment of non-participating firms. However, such an approach would lead to an overestimation of the policy's treatment effect for the following reasons:

- Programme participation is the result of a self-selection process. In order to participate, firms have to apply with a project proposal and, thus, firms self-select into the programme. It could be argued that applying firms would have possibly conducted more R&D, even without the programme participation, compared to a random firm in the economy because an application requires having (good) R&D project ideas. Moreover, the firm might also implement some of them without public support; therefore, on average, the group of applicant firms conducts more R&D than non-applicants, even in absence of the policy.
- Receiving an R&D grant, conditional on the application, depends on a further selection by the management authority. It seems plausible that firms that have some track record of successful R&D, and a sufficiently high level of ongoing R&D, are more likely to be selected than others with lower values in these factors because public agents want to maximise programme success. In short, the agency's selection might also favour firms that would conduct more R&D than a random firm in the economy, even in absence of the policy.
- The applicant firms may also have other structural characteristics that facilitate R&D, even in the absence of the policy, such as greater internal financial resources, higher human capital and so forth.

For these reasons, it is impossible to estimate the counterfactual situation ("what would the firms have invested in R&D had they not been supported?", $E(Y^C|S=1)$) simply as the average R&D spending of non-participating firms. However, two intuitions on how to define such an estimate of the counterfactual describe the essence of the two most commonly used econometric evaluation techniques.

The first intuition an evaluator could have is the one that makes the sample of non-participating firms more comparable to the participating firms, rather than just using a sample of randomly chosen non-participants as a control group. This is the basis for the so-called 'matching estimators.'

A second intuition that an evaluator could have involves using prior observations about how much the participating firms have spent on R&D (i.e., prior to their entry into the programme). This intuition forms the starting point for the most commonly used evaluation technique used at present, the 'difference-in-difference (DiD) estimator'. One compares *changes* in R&D inputs (or other performance variables) between participants and non-participants in the DiD estimation.

2.1 Matching

2.1.1 General intuition

Matching estimators have been broadly applied and discussed throughout the economic literature (see Smith and Todd, 2005, for an econometric overview on Matching estimators; and Zunica et al., 2014, for an overview on applications in the context of R&D subsidies).

In the case of matching, the counterfactual situation of treated firms (i.e., the "untreated outcome") is constructed from a control group of firms that did not receive innovation subsidies. The matching relies on the intuitively attractive idea of forming two comparable groups by balancing the observable characteristics of programme participants with those of a comparable group of untreated firms. Remaining differences in the outcome variable between both groups are then attributed to the treatment (i.e., to the programme participation) because differences in other observable characteristics are taken into account.

The counterfactual situation cannot simply be estimated as the average outcome of any sample of non-participants because of a potential selection bias, due to the fact that the receipt of a subsidy is not randomly assigned (i.e., $E(Y^C|S=1) \neq E(Y^C|S=0)$). Instead, one assumes that firms with very similar observable characteristics behave similarly to the supported firms. This assumption implies that they could have

participated in the programme with the same likelihood as the participants, but did not do so because of statistically random (i.e., not systematic) reasons.[4]

Examples of such observable characteristics, which have been used to balance the samples of participants and selected non-participants in the case of Research, Development and Innovation (RDI) support programmes, include firm size and age, the economic sector and variables that describe the firms' experience with public support programmes as well as previous (successful) R&D projects that could trigger further investment.

The critical point around the assumption underlying the matching approach is whether or not we can really observe all factors that determine selection into the programme. This assumption cannot be tested and it remains up to the evaluator who applies this technique to find convincing arguments that the observed characteristics plausibly allow for the assumption to be satisfied. If this is the case, then the unobserved counterfactual situation (i.e., participating firms that are hypothetically observed when not receiving the treatment, $E(Y^C|S=1)$) can be approximated by, for example, the average R&D spending of a selected group of non-participating firms that are on average equal in relevant observable characteristics.

2.1.2 Matching approaches

One of the most commonly used matching estimators is the 'nearest neighbour (NN) matching estimator'. The NN matching estimator pairs each subsidy recipient (treated) with the single closest non-recipient (control) in terms of observable characteristics.

NN matching does not necessarily have to be limited to a single nearest neighbour. Suppose one has a sample of 200 programme participants and a very large number of possible control observations, (2,000,000 non-treated firms, for example). In this case it might appear tempting to choose two nearest neighbours for each subsidised company as comparison units. One could also choose more than two neighbours. When doing so, however, the evaluator faces a common trade-off that is present in many econometric settings: the estimation's bias versus efficiency. For reasons of efficiency of estimators, it is always preferred to use large samples because the estimate's precision increases in larger samples. However, the more neighbours one chooses for the control group, the more dissimilar each further neighbour that is being picked will become. The more dissimilar the neighbours become, the more biased the eventual estimate of a treatment effect may become.

While the optimal number of neighbours cannot be determined a-priori, scholars have developed several variants of matching (cf. Todd and Smith, 2005, for an overview). In order to avoid bias, one can pre-define an acceptable neighbourhood of similarity around the chosen characteristics (e.g., firm sizes that are smaller or larger than the treated firm within a chosen interval). If no neighbour can be found within that interval, then the treated firm is excluded from the matching estimate. This is called 'calliper matching'. One can also define a neighbourhood and keep all firms that are within the chosen interval as neighbours in order to use multiple neighbours for efficiency reasons according to the estimators. This is called 'radius matching'. In an extreme case, one can use all non-subsidised firms to form a counterfactual for each treated firm. This is called 'kernel matching' and the neighbours are then weighted by their similarity, such that their weights add up to unity. Kernel matching estimators are the most efficient version of matching, but they will always be more biased than the NN matching with a single neighbour. The latter is, in contrast, the least efficient in the class of matching estimators. As a rule of thumb, one can conclude the following: if the application of the NN matching with a single neighbour already leads to statistically significant treatment effects, then the analysis is concluded. If these treatment effects are not statistically significant, then scholars may turn to other variants of matching estimators that lead to more precise estimates (i.e., higher statistical efficiency). It is common practice to use more than one method in applied work in order to demonstrate how robust the results are or, in other words, how sensitive the results are to the choice of various estimation techniques.

2.1.3 Implementation

In order to implement a matching estimator, one must first consider which variables should be taken into account. This question cannot be answered in a general form, but the chosen variables should determine the treatment and should depend on neither the treatment nor the outcome variable. The variables have to be selected carefully on a case-by-case reasoning because each programme might have specific participation rules and specific purposes. One should keep in mind that the variables must be "exogeneous" (i.e., are unaffected by

[4] This assumption is called the conditional independence assumption (CIA) and it implies that programme participation and potential outcome are statistically independent for firms with the same set of exogenous characteristics (Rubin, 1977).

participation in the programme). Ideally, one chooses variables that are measured prior to programme participation. In addition, they should not be affected by the anticipation of the treatment.[5]

Examples of variables that are often used in studies on R&D grant programme participation are: firm size (measured as numbers of employees, sales or total assets), firm age, the economic sector, internal financial resources, debt levels, the capital intensity of the firm, human resources, experience with the public funding system (e.g., former project applications) and evidence of a previous track record of successful innovation (e.g, patent applications).

The number of characteristics on which one would like to match the treatment and control group often become quite sizable; for example, the set of variables often exceeds 10. Furthermore, the variables have different dimensions. Firm size might be a headcount of the number of employees, but the sector of activity is described by a set of binary indicators labelling the food industry, the automotive sector and so forth, and age is measured in years. In that case, one would have to find the neighbours by calculating a multidimensional distance that can determine which firm is the closest neighbour.[6] Even though calculating such multidimensional distances is less problematic with modern computing power, scholars have referred to this as "curse of dimensionality" and have circumvented such (previously cumbersome) matrix calculations by using a statistical insight: it is sufficient to match on a single index, the so-called 'propensity score' (Rosenbaum and Rubin, 1983) in order to balance the treatment and control groups according to certain covariates.

In this case, scholars refer to the implementation as *propensity score matching*. The propensity score is the estimated probability of programme participation. This can be computed for each firm in the sample by estimating a regression model in which the participation (yes/no) is determined by the firm characteristics chosen.[7]

It is essential that there is enough overlap between the control and the treated group regarding their participation probabilities for a successful application of the matching estimator. This condition is usually referred to as "common support." For example, some of the participants might have attributes that almost certainly determine participation; therefore, they have an estimated propensity score (participation probability) of 0.9 and above, but no control has a probability of 0.9 (or nearby); thus, this firm has no comparison group and it is outside the common support area. Other participants might not have such deterministic attributes, and some might have even entered the programme even though their predicted participation probability is rather low (e.g., 0.15). The common support assumption implies that the possible control group must contain comparable firms across all of the treatment group's estimated propensity score such that "close neighbours" can be found. In practice, the treatment group and control group are often restricted to common support. One could calculate the minimum and the maximum of the potential control group's propensity scores and delete observations on treated firms with probabilities larger than the maximum, and smaller than the minimum in the potential control group. The participants that are dropped from the analysis are usually a very small share of the sample's total number of observations. Strictly speaking, the restriction to common support does not lead to an estimate of the treatment effect on the treated, but of a "treatment effect on the treated in the area of common support".

Once the samples have been matched, according to the participation probabilities as determined by the firm attributes, the researcher can compare the R&D spending of programme participants and their matched controls. Common two-sample t-tests may be used in the most simplistic settings in order to conduct statistical hypothesis tests.

---

[5] Ashenfelter (1978) was the first who documented that future participants strategically reduced their wage before the programme application in order to become eligible in the case of a labour market programme where the participation eligibility was defined by a wage below a certain threshold. This would rule out "wage just before the programme participation" as a valid matching criterion.

[6] An example of such a multidimensional distance is the Mahalanobis distance (*MD*) $MD_{ij} = \left(Z_j - Z_i\right)' \Omega^{-1} \left(Z_j - Z_i\right)$ which can be calculated between each treated firm *i* and all possible control firms *j* in the sample. *Z* here refers to the firm's characteristics that were chosen to be taken into account and Ω is the empirical covariance matrix of these matching arguments, based on the sample of potential controls.

[7] The regression models are usually probit or logit models.

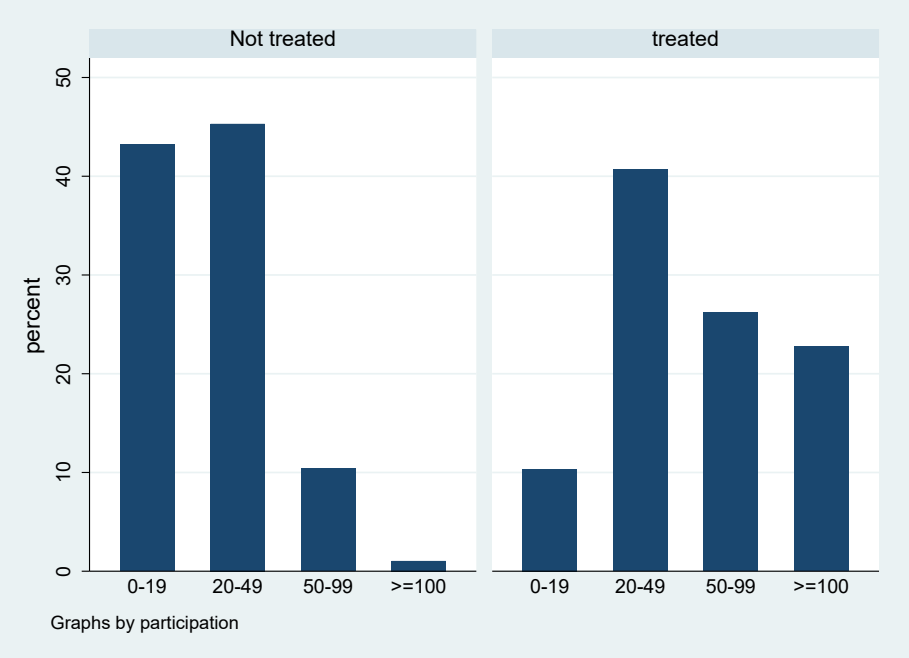
2.1.4 A hypothetical example using simulated data

In order to illustrate the matching method, let us consider the following hypothetical example: an R&D grant programme for innovative, small and medium-sized firms (SMEs) had 145 participants; in total, data for a sample of 1,200 SMEs has been collected, including 145 participants and 1,055 non-participants.

A standard measure in the field of the economics of innovation is considered as outcome variable: the firms' R&D intensity in percent is considered (R&D expenditure in € divided by sales in € times 100). When looking at the average R&D intensity in the two groups, the treated firms and non-treated firms, we find that the average R&D intensity in the participant group amounts to 6.0% and amounts to 5.45% in the group of non-participants. Thus, the effect of the programme at first glance could be 6.0% – 5.45% = 0.55 percentage points in R&D intensity that is triggered by the programme.

When having a look at some of the firms' structural characteristics, however, one also finds that the participating companies are, on average, larger than the non-participating companies (see Figure 1). It can be seen that the smallest firm size category of 0-19 employees includes more than 40% of the firms in the potential control group of non-participants, but this share only amounts to about 10% in the treatment group. In contrast, more than 20% of the participants have over 100 employees, whereas the share of such firms is only about 1% in the not-treated firms. This is a common pattern observed in reality, even if the programme is targeted at SMEs. Even among SMEs, it is typically the case that larger firms have a more professional R&D management or general administrative unit, and, therefore, may be more informed about government programmes and also have more project ideas at a given point in time (i.e., they are more likely to apply).

Figure 1: Hypothetical firm size distribution among R&D programme participants and non-participants[1]

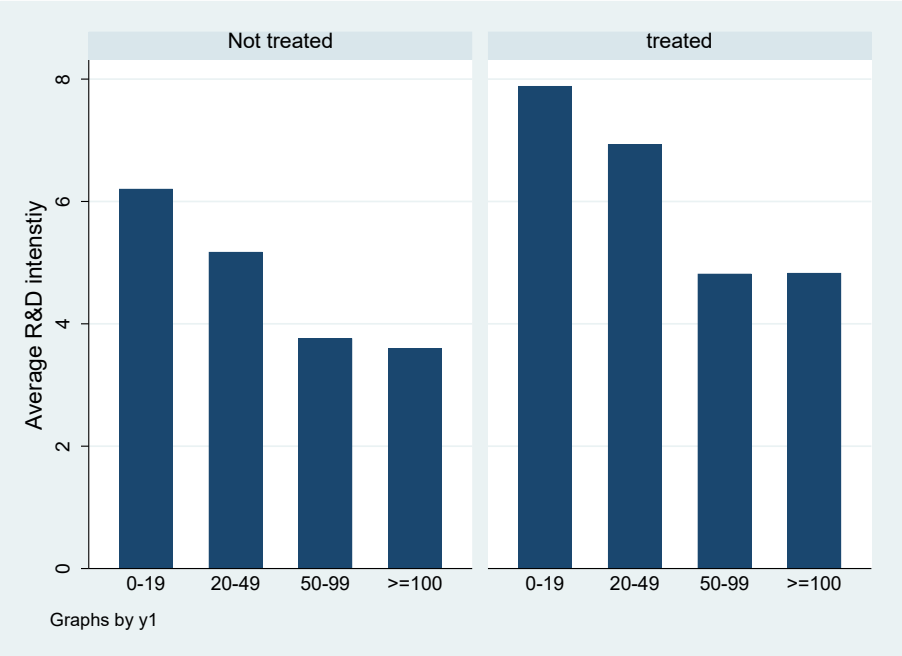


[1]*Firm size is the number of employees*
Source: Author's simulated data.

If one looks at the outcome variable, R&D intensity and how this relates to firm size, it can be seen that the average R&D intensity of firms decreases with firm size in both the treatment group and the possible control group (see Figure 2). The smallest firms in the treatment group achieve an R&D intensity of almost 8%, but the value declines to about 7% in the group of firms that have between 20 and 49 employees and drops further to about 4.5% in both groups of larger firms that have over 50 and over 100 employees. In the non-treated firms' sample, the R&D intensity amounts to slightly more than 6% in the smallest firms, but drops to values below 4% in the larger firms' categories. This pattern is commonly observed in reality. As firms become larger, they

tend to have more established product portfolios that lead to higher sales. If R&D investment is put in relation to sales, then it is typical that the R&D intensity will be higher among the smaller firms.

Figure 2: Hypothetical average R&D intensity by firm size and programme participation[1]



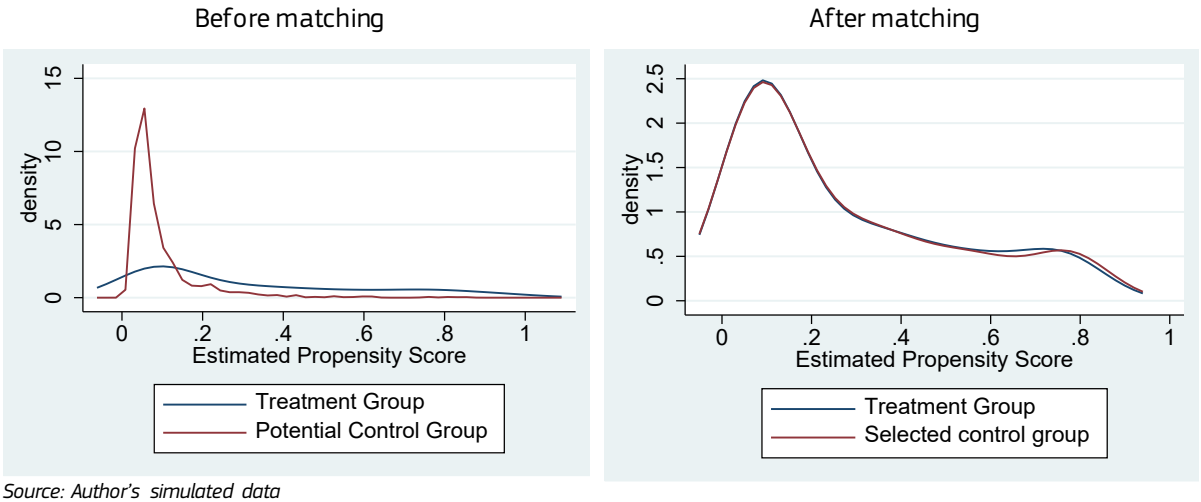
[1]*Firm size is the number of employees*
Source: Author's simulated data.

These relationships between variables invalidate the unadjusted comparison of average R&D intensities between programme participants and non-participants. On the one hand, programme participation increases with firm size; however, the outcome variable R&D intensity decreases with firm size on the other hand.

Matching is a method that can eliminate such structural differences between the treatment and control groups. In this over-simplified example, one could directly match the participating firms to "twins" with the same firm size from the non-participating firms. Alternatively, one could determine the propensity score by estimating a regression equation with the participation indicator as the outcome variable and firm size as an explanatory variable. After doing so, one finds that the average participation probability amounts to 31% in the treatment group and it ranges from 3% to 100%. The average is only 9.5% in the potential control group and the probability ranges between 2.5% and 86%. The whole distribution of estimated participation probabilities can be seen in Figure 3. The figure also shows the distribution of participation probabilities after the participating firms have been matched to "twins" of the same size in the control group. After the matching, it can be seen that the distributions are almost exactly on top of each other, except at high values of participation probabilities beyond 60%. In these areas, it turned out to be more difficult to find firms with very similar sizes in the control group.

When running the NN-matching algorithm, it also turned out that 8 observations from the participant group were dropped because no suitable neighbour could be found for them (i.e., they were off the common support range).

Figure 3: Estimated distribution of participation probability before and after matching



Source: Author's simulated data

One finds the following results after the sample of participants has been matched to twins: 137 participants out of the 145 could be matched to non-participating firms that had a very similar participation probability; in this case, this coincided with having a very similar size. After the matching has been executed, the average firm size in the participant group is 59.2 employees and is 59.0 in the selected control group. The average participation probabilities amount to 27.5% and 27.4% respectively.

The R&D intensity in the group of participants amounts to 5.96% on average. This number is 4.36% in the matched control group. The latter is the estimate for the counterfactual (i.e., what the participating firms would have invested had they not been subsidised). This implies that the policy's causal effect (under the validity of the model assumptions, and under the conditional independence assumption specifically), is 5.96% - 4.36% = 1.6 percentage points. One should recall that, prior to the matching of the samples, this difference only amounted to 0.55 and would, thus, have led to a misleading underestimate of the policy's efficacy.

In reality, of course, the matching problems are more complex than illustrated here, as more variables than just firm size affect participation probabilities and programme outcomes. Therefore, some examples of matching studies using real data are briefly presented below.

2.1.5 Examples of matching studies in the RDI context

The first study that applied the matching estimator in the RDI context is Almus and Czarnitzki (2003) who investigated the effect of R&D subsidies on innovation activities in Eastern Germany after the re-unification in the 1990s. They found that total crowding-out effects can be rejected and that, because of the subsidy, firms increased their innovation activities by four percentage points relative to a counterfactual scenario of not receiving the subsidy.

As the Almus/Czarnitzki-study investigates the special situation of Eastern Germany in the transformation process from a planned economy to a market economy, its results might be not representative for counterfactual impact evaluation in other EU countries.

A potentially more representative study for RDI programmes in other Member States is Czarnitzki and Lopes-Bento (2013). They investigated a Belgian R&D programme and studied its causal effect on firms' R&D employment. They utilised a sample of 292 subsidised firms and 4,469 potential control firms. On average, the subsidised firms had about 18 R&D employees and the non-subsidised firms had about 2.6 employees, on average.

However, when comparing the subsidised and non-subsidised firms in terms of observable characteristics, it turned out that the two groups differ systematically: among other attributes, the subsidised firms were on average larger in terms of employment, were more export-active, had previously filed more patents for their inventions and were also more active in applying for subsidies in the pre-treatment period.

Therefore, Czarnitzki and Lopes-Bento applied propensity score matching and drew one nearest neighbour for each subsidised company. After the matching, the comparable control group had about 8 R&D employees on average, not 2.6 as mentioned before the matching was conducted.

They derived total programme effects by conducting a back-of-the-envelope calculation using their estimates and detailed information on total programme budgets. According to their calculations, the programme created 16,800 R&D employment person-years with a budget of EUR 628 million and resulted in EUR 37,000 per created R&D-person year as societal cost.

They also investigated whether the programme effects were heterogeneous across some dimensions of interest. One could be concerned that firms show heterogeneous treatment effects if they were receiving:

(i) multiple grants for different projects at the same time;

(ii) different grants consecutively in different funding rounds;

(iii) grants from other programmes.

Such heterogeneity analyses can be done by analysing the respective subsamples of programme participants separately. Czarnitzki and Lopes-Bento measured these cases by regressing the estimated treatment effects on variables. They did not find any evidence for such concerns. The estimated treatment effects were not significantly different (i.e., smaller for any of these subsamples of firms).

2.2 Difference-in-differences (DiD)

The most common method that is used in the context of counterfactual impact evaluations including RDI programmes nowadays are DiD approaches. However, while matching estimators could be applied with cross-sectional data (i.e., data from a single point in time), the application of DiD techniques requires panel data.

The treatment group's outcome variables of interest are observed prior to programme participation and thereafter. The control group's outcome variables are observed for the same time period. The idea of the difference-in-difference (DiD) estimator is based on exploiting this "panel structure" (i.e., different firms can be observed over time). The DiD estimator works as follows: one could calculate the difference in outcomes, *Y*, for each observed firm over time (i.e., for both the treated firms and for the control group). Let us suppose that period t_1 is the treatment period and that t_0 is a year prior to programme participation:

$$\Delta_i^T = Y_{i,t1} - Y_{i,t0}$$

$$\Delta_j^C = Y_{j,t1} - Y_{j,t0}$$

where *T* denotes the treatment group, *C* the control group and *Y* is an outcome variable of interest, such as R&D employment or expenditure. One can calculate the change of *Y* over time. However, the change in *Y* may also be caused by other economic factors that are drive the change in Y for both the treatment and the control groups. Hence, an underlying assumption of this method is that such 'external' factors affect both groups in the same manner. Thus, the treatment effect α can be estimated as difference between both differences (i.e., pre-post difference in the outcome for the treated minus the pre-post difference for the non-treated firms):

$$\alpha^{DiD} = E(\Delta_i^T) - E(\Delta_j^C).$$

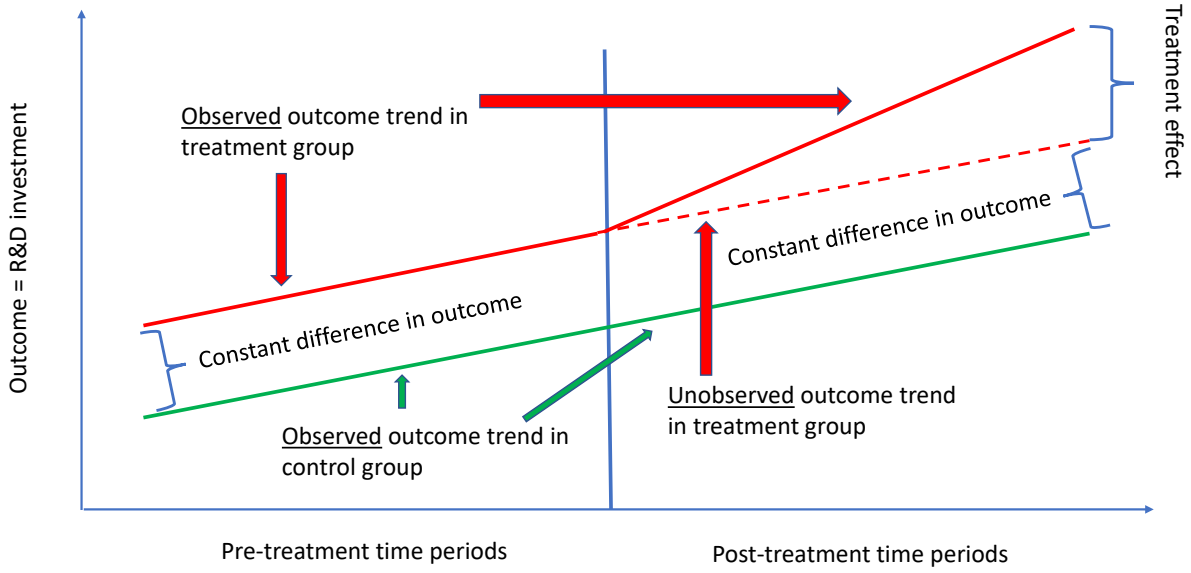
The expected value would simply be estimated as the sample averages of the changes in *Y* in the treatment and control group respectively.

It is possible to show that a test of whether or not the treatment effect is positive in statistical terms (the programme increases *Y* in the funded firms) could be implemented by a standard two-sample t-test on mean differences in this example.

In practice, scholars often use more than two periods to estimate treatment effects with DiD methods. Figure 4 shows a stylised graph of how a programme's effects can be derived. One can compare how the trend in an outcome variable of interest (e.g., R&D investment) develops over time between a treatment group of programme participants and a control group of non-participants. The green line is the trend in R&D investment for the control group and the red line is R&D investment in the treatment group. The two components for deriving the treatment effects are the two differences between the red and green lines before any firm received the treatment and the difference between the two lines after the programme participants received the treatment. The figure shows that one estimates the treatment effect implicitly by assuming that the trend that

was observed prior to the treatment would have continued, even if the participants would not have been in the programme. This counterfactual is depicted by the dashed red line.

Figure 4: Stylised DiD methodology



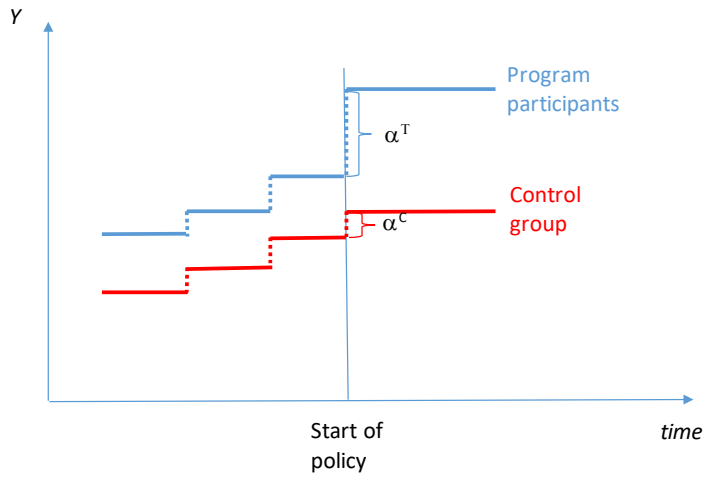
One major advantage of observing both treated firms and non-treated firms over time is the fact that the method allows for controlling heterogeneity (i.e., time-constant, but unobserved to the researcher). For example, the treated firms in Figure 4 always invest more into R&D than the firms in the control group (i.e., the red line is always above the green line). This could be due to unobservable differences among the groups; for instance, the programme participants could have a better R&D management that delivers better ideas for research projects relative to the control firms. In that case, it is plausible that these firms would always invest more than the control group. In the DiD methodology, the unobserved attributes that might influence the outcome variable are not problematic as long as these are time-constant, because the treatment effect is not derived from differences between the groups at some point in time, but changes in investments over time.

2.2.1 The common trend assumption

The fact that the red and green lines in the figure above follow the same trend in the pre-treatment time periods is a necessary condition in order to consistently identify programme effects. If the trends were not common, then one would not believe that the control group would have behaved in a similar way as the treatment group after the programme participation, independently of the fact that they did not participate in the programme. In that situation, the comparison of the difference in the differences prior to and after the treatment would not lead to meaningful results. This required a condition of parallel trends before the treatment period and is known as the "common trend assumption".

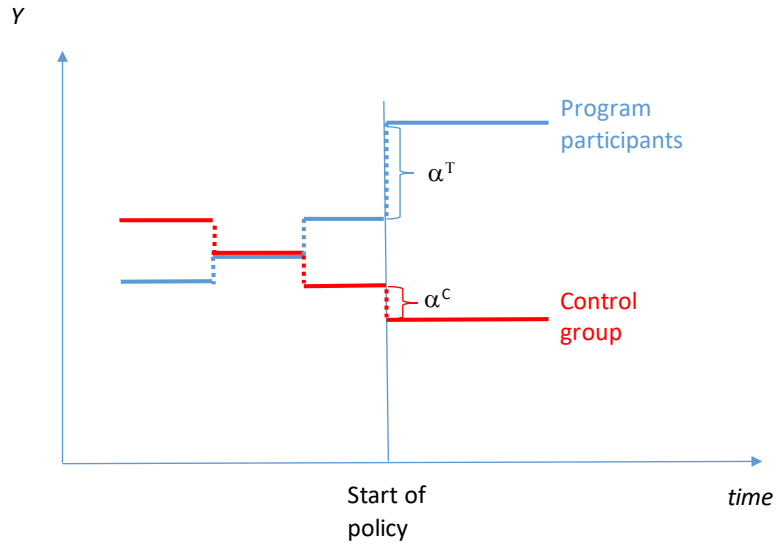
In Figure 5, hypothetical data show that the programme participants and the control group evolve in the same way over time. However, when the participants enter the programme, their increase in the dependent variable Y (e.g., R&D investment) is larger than that of the control group. The estimated treatment effect α^{DiD} would amount to $\alpha^{DiD} = \alpha^{T} - \alpha^{C}$. The treatment effect would not be α^{T} because this would only be a before-after comparison of the participants. As can be seen in the graph, the investment develops positively, irrespectively of the programme and this can be seen in the positive trend in Y of the control group. Therefore, subtracting α^{C} from α^{T} adjusts the estimated treatment effect by netting out the general economic development (i.e., the possible positive effect that would have also occurred in absence of the policy), thereby isolating the programme's true effect under the assumption that both the treatment group and the control group would have evolved similarly independently of the policy. This assumption becomes credible because of the common trend that can be observed before the policy had been in place.

Figure 5: A sketch of the DiD methodology with a common trend



However, if the common trend assumption is violated, then the estimated treatment effect obtained by a difference-in-difference approach would not typically be regarded as trustworthy. Such a situation is depicted in Figure 6 below. In this graph, the two groups of firms never exhibit the same trend and, therefore, the application of the difference-in-difference methodology would mistakenly result in a treatment effect that is equal $a^{DiD} = a^T - a^C$, but in this case it seems not to be credible. In fact, $a^{DiD} > a^T$ as the treatment effect even gets boosted by the negative trend of α^C. The graph suggests that the two groups are not comparable at all and, thus, using the control group in a standard difference-in-difference framework is an invalid approach in this case. It can be observed from the pre-treatment trends that the outcome variable *Y* would have increased anyways even in the absence of the policy, and that *Y* in the control group would have decreased even in the absence of the policy. Therefore, this control group does not yield a valid comparison.

Figure 6: A sketch of the DiD methodology with a violation of the common trend assumption



Box 2. Difference-in-Difference (DiD)

DiD approaches have become the most commonly used techniques for treatment effects estimation.

Advantage:

Unlike matching, DiD approaches also account for selection on unobservable attributes of programme participants, as long as these are time-constant in the period under review (e.g., R&D management quality).

Disadvantages:
Panel data are required for their application;
Estimation rests on some strong assumptions, such as common trends of treatment and control group in pre-treatment time periods.

2.2.2 Conditional difference-in-differences

A possible solution to the violation of the common trend assumption in the context of difference-in-difference is the combination of the methodology with matching estimators. In very simple terms, this means that the control group would not use all of the firms that did not participate in the programme, only firms that are similar to the participant groups in some observable characteristics. For instance, the control group could be adjusted by discarding firms that are not in the same industries and regions as the treatment group. In addition, other firm level variables could be considered. The choice should be made in an economically meaningful way for each application. In practice, scholars often have to revert to conditioning on similar pre-treatment growth rates between control and treatment group, as the treatment group often appears to be a very specific type of firm (e.g., firms with a high desire for fast and large growth that, consequently, conduct many R&D projects). One often has to rely on calculating pre-treatment growth rates and to pick the control group based on these because a "growth intention" is unobserved to every researcher who is interested in deriving treatment effects. The common trends across the two groups can then be restored and, therefore, a policy treatment effect can (credibly) be derived.

In summary, the Conditional Difference-in-Difference (CDiD) methodology can be seen as an attempt to move from a situation as sketched in Figure 6 (i.e., a violation of the common trend assumption) to a scenario as drawn in Figure 5, a common trend across the treatment and control groups.

2.2.3 An example of the EU's Seventh Framework Programme

Szücs (2018) evaluates the impact of the 7th Framework Programme for Research (FP7)[8] on innovation activities of subsidised firms. In the study, the author is specifically interested in evaluating the possible benefits of collaborative research, especially between firms and public research institutions such as universities. Research collaboration among multiple partners, who apply for funding as a consortium, is often encouraged in public R&D funding schemes.

The author identifies FP7 participants from the CORDIS[9] database and links them to the Orbis firm-level database which contains accounting data. He also collects patent information from the EPO PATSTAT database.[10] A sample of about 2,600 FP7-participating companies is then matched in pre-participation periods to a control group of firms that is selected from about 44,000 firms from the Orbis database. After the matching procedure, a panel of 2,187 treated-control firm-pairs is retained. Among other results, Szücs (2018) finds that FP7 participants patent more after participation in FP7 compared to their prior activity and relative to the evolvement of the patenting of the matched control group (i.e., the CDiD-results find a positive treatment effect with regard to inventive activity).

Furthermore, he investigates whether the size of the consortia has an impact on the treatment effects and whether the involvement of universities and public research centres increases the programme's effects. He concludes that the larger the consortia, the more within-cooperation spillovers can be obtained; therefore, the treatment effects are an increasing function of the number of consortium members. Cooperating with universities, especially with well-ranked ones in international comparisons, further increases the FP7 effects on participating firms.

[8] The 7th Framework Programme was the EU's research funding programme between 2007 and 2013.

[9] The Community Research and Development Information Service (CORDIS) is the European Commission's primary source of results from the projects funded by the EU's framework programmes for research and innovation, from FP1 to Horizon Europe.

[10] A description of these data sources, together with the other sources most commonly used for evaluations of R&D grants, is presented in subsection 3.1.3.

Methodologically, Szücs conducts a CDiD[11] study that accounts for unobserved effects at the country and industry level, but not at the firm level. This implies that unobserved firm-level heterogeneity among participants and non-participants is only accounted for at the average level in the DiD model, but are not firm-specific. The latter is a common standard in current literature and, therefore, is a limitation of this (otherwise highly interesting) study.

2.3 Regression Discontinuity Design (RDD)

The RDD relies on the idea of exploiting programme characteristics for the evaluation because these are exogenously determined by the policy makers and, by assumption, cannot be influenced by potential participants. A common example is the case of policy schemes in which participants are chosen based on a peer-review process in which experts assign scores on the basis of project quality to the participants.

An example is the Eurostars program, which is a joint programming initiative run by EUREKA, a research network of European countries that aims to stimulate close-to-the-market R&D for civilian purposes.[12] Eurostars was launched in 2008 to provide financial support for innovation activities in R&D-performing SMEs.

The international consortia of SME can file applications for RDI subsidies. After each application round, multiple experts assign scores between 0 and 600 to each application (see e.g., Hünermund and Czarnitzki, 2019). A minimum score of 400 is required to qualify for funding.

The framework of the Eurostars programme is almost a textbook application of a so-called Regression Discontinuity Design (RDD). In such a setting, firms are selected into the programme because of the score obtained. Econometricians distinguish between two variants of RDD, a sharp and a fuzzy RDD, within the context of an RDD.

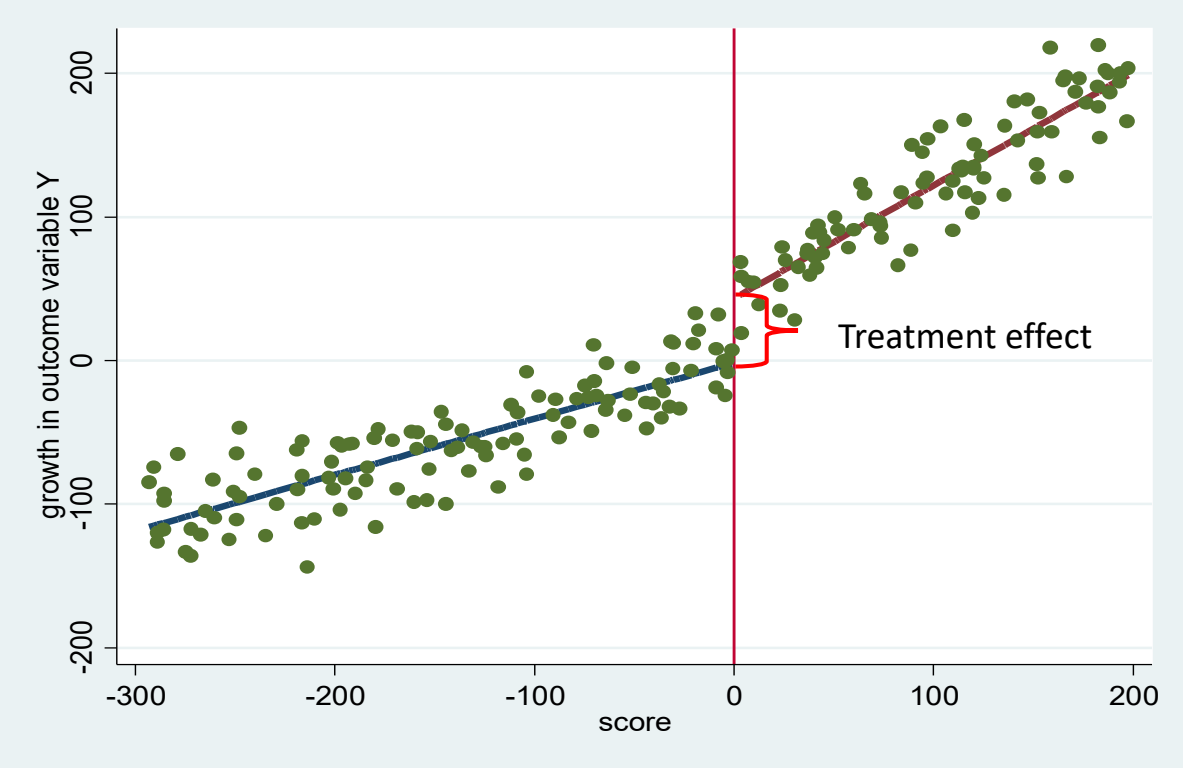
2.3.1 A hypothetical example for Sharp RDD

The idea of a sharp RDD is described for purely illustrative purpose in order to explain the logic of RDD. In the Eurostars programme, firms would be funded if they pass the threshold of 400 in the scoring model of the peer-review process; otherwise, the submitted proposals would not be retained for funding. This means that the subsidy would not be given randomly to firms, but would be given strictly based on the scores. In such cases, one could typically expect a relationship between a certain target variable (e.g., growth in R&D spending) and the scores as depicted hypothetically in Figure 7. In the figure, the score has been normalised by subtracting the value of 400, such that the threshold value is now 0 and positive values indicate being above the threshold score, whereas negative values indicate being below the threshold score.

[11] Other recent examples of studies employing conditional DiD methods in the context of RDI programmes are Fornaro et al. (2020) or Novadays (2020).

[12] Cf. https://www.eurekanetwork.org/

Figure 7: Hypothetical illustration of data conforming to a sharp RDD



The underlying idea in the simulated data is as follows: if one were to simply compare the mean of the growth in the outcome variable, then one would find that the values are -57 for non-funded firms (i.e., observation below the threshold, and 123 above the threshold, that is, for funded firms). Thus, the DiD estimator would estimate a treatment effect of 123 – (-57) = 180.

This would, however, be misleading because the graph illustrates that firms which got funding differ in two ways from the other firms: first, their observations are shifted upwards because of the treatment (i.e., higher growth in Y); second, when looking at the plot, it also shows that the relationship between outcome variable and score shows a steeper slope for the score in the group of funded firms. This reflects the possibility that the firms that obtain better scores are simply the "better" firms on all accounts (i.e., they might have grown more than the firms without a subsidy even had they not received the treatment).

Instead, the RDD estimate relies on the idea that firms ranked closely around the threshold score have very comparable projects in terms of their quality. By comparing firms closely around the threshold, one tries to approximate a situation where the treatment assignment has happened quasi "randomly" and not based significant structural or quality differences of the submitted proposals.

The correct treatment effect could be estimated here in this example by estimating a regression equation with growth in Y as outcome variable and, as explanatory variables, a treatment indicator (i.e., a dummy variable taking the value 1 if a firm is ranked above the quality threshold, and 0 otherwise) and the score which is interacted with the treatment dummy. This estimation allows obtaining the treatment effect of interest that can, ,be identified visually, in the graph, by the vertical difference between the two regressions lines at the quality threshold, that is, the discontinuity point created by the subsidy. Thus, the RDD estimator would identify a "local" treatment effect around the threshold. In this illustration, the simulated treatment effect actually amounts to 50 units of higher growth in Y, and not 180 as the DiD estimator would suggest.

Note that the RDD estimator thus identifies a local treatment effect around the threshold, whereas the DiD estimator identifies an average treatment effect for all treated observations. This is a major disadvantage of the RDD estimator. While it has a strong internal consistency, it can only determine the treatment effect at the threshold locally, but this is often not of interest for the managing authorities. Usually, one is not particularly interested in the effect in firms that have just made it into the programme. An average including the best

participants seems more informative for concluding whether a certain policy programme is successful and whether it should be continued or adapted.

2.3.2 An example using Sharp RDD

Bronzini and Iachini (2014) use a sharp RDD in order to estimate the effects of a subsidy programme in Northern Italy. The submitted research proposals were examined by an independent technical committee that assigned scores to the proposals. Only those above a critical threshold were subsidized. The maximum score amounted to 100 and the required points to cross the funding threshold were 75. As outcome variable Bronzini and Iachini (2014) employ tangible and intangible assets obtained from balance sheet data, as they had no information on R&D and other innovation-related variables.

The authors estimated a regression equation with the outcome variable on the left hand side and the treatment variable and flexible functional forms of the assigned scores as explanatory variables. They also vary the neighbourhood around the threshold value of 75 which results in sample sizes of 357, 171 and 115 firms, respectively. This is often done in practice to investigate the robustness of the results, as there is no natural choice for the "bandwidth", i.e., the neighbourhood around the threshold (more details on RDD estimation, including on functional form and bandwidth around the threshold, are presented in subsection 2.3.4).

They cannot reject the hypothesis of full crowding out effects for this program, i.e., they cannot confirm that tangible or intangible assets increase as a response to the subsidy. In this case, the programme would not conform to the EU State aid rules as it is not effective, but tax money is used to finance it. It would thus be an inefficient allocation of public funds. However, when differentiating the treatment effect by the median of firm size in their sample, they find that smaller firms show statistically significant positive treatment effects, i.e, the programme participation causes increased investment into assets. For larger firms, no effects are found.

2.3.3 RDD applied to fuzzy settings

The example above was used to introduce the logic of Sharp RDD estimation. In some cases, however, the discontinuity is not sharp but fuzzy instead. An example is the Eurostars programme in which highly innovative small companies can apply as international consortia for funding projects that are close to the market. The projects are evaluated on a scale between 0 and 600 during a peer review process and the minimum threshold for funding is 400. In practice, however, there are also proposals that were evaluated above 400 points, but still do not receive funding. The reason is that the national managing authorities have set earmarked budgets before the funding rounds, and if these are exhausted because there were higher ranked proposals that obtained funding, it happens that even submitted project above the eligibility threshold are not supported (see e.g., Hünermund and Czarnitzki, 2019, for more information).

This creates the situation of a "fuzzy" RDD. Here the funding is not strictly a function of the score, but also happens because of some further, often unobserved, factors. In the fuzzy RDD framework, one cannot only make use of the difference between funded and non-funded firms at the discontinuity point on order to derive an estimate of the treatment effect, but can in addition exploit the fact that observations of firms that passed the quality threshold but did not get funding are available. These firms might be the "better" firms when compared to the non-funded firms in terms of unobserved characteristics such as management quality, R&D capabilities, and so forth, but it could be assumed that they are similar, in these unobserved qualities, to the funded firms. Therefore, one can make use of these additional observations in order to derive a treatment effect. Estimating the treatment effect is slightly more complicated compared to the sharp RDD. In the framework of a fuzzy RDD, an instrumental variable (IV) (more details of this method are presented in the next section) regression is estimated, where the treatment indicator is instrumented with a binary indicator of whether a submitted proposal has been ranked above or below the quality threshold by taking into account the scores as in the example above. The estimator can be implemented with the familiar Two Stage Least Squares (2SLS) approach. In the first stage, the funding (treatment) indicator is regressed onto the above/below-threshold indicator and the scores of funded and non-funded firms (and possibly other covariates included in the model). The linear prediction obtained from this regression is used in a second stage instead of the actual treatment dummy.

Box 3. Regression Discontinuity Design (RDD)

Advantages:

RDD is often regarded as the most internally consistent estimation, that is, it most credibly establishes causality of the estimated effects, because it utilises some exogenous threshold imposed by the programme to form treatment and control groups, and evaluates the programme effects around that threshold. It is plausibly assumed that firms around the threshold, e.g., just below and above eligible firm size, behave very similar besides one group receiving the treatment.

Disadvantage:

Often not applicable, as the design of RDI programmes may not offer the opportunity to define a required threshold value at which programme effects could be evaluated.

A reasonably large number of data points are needed closely around the threshold, e.g., many firms that are just below or above a certain firm size.

Cannot be used to estimate overall programme effects, such as average treatment effects for the total group of participants, as it only estimates effects "locally" around the threshold.

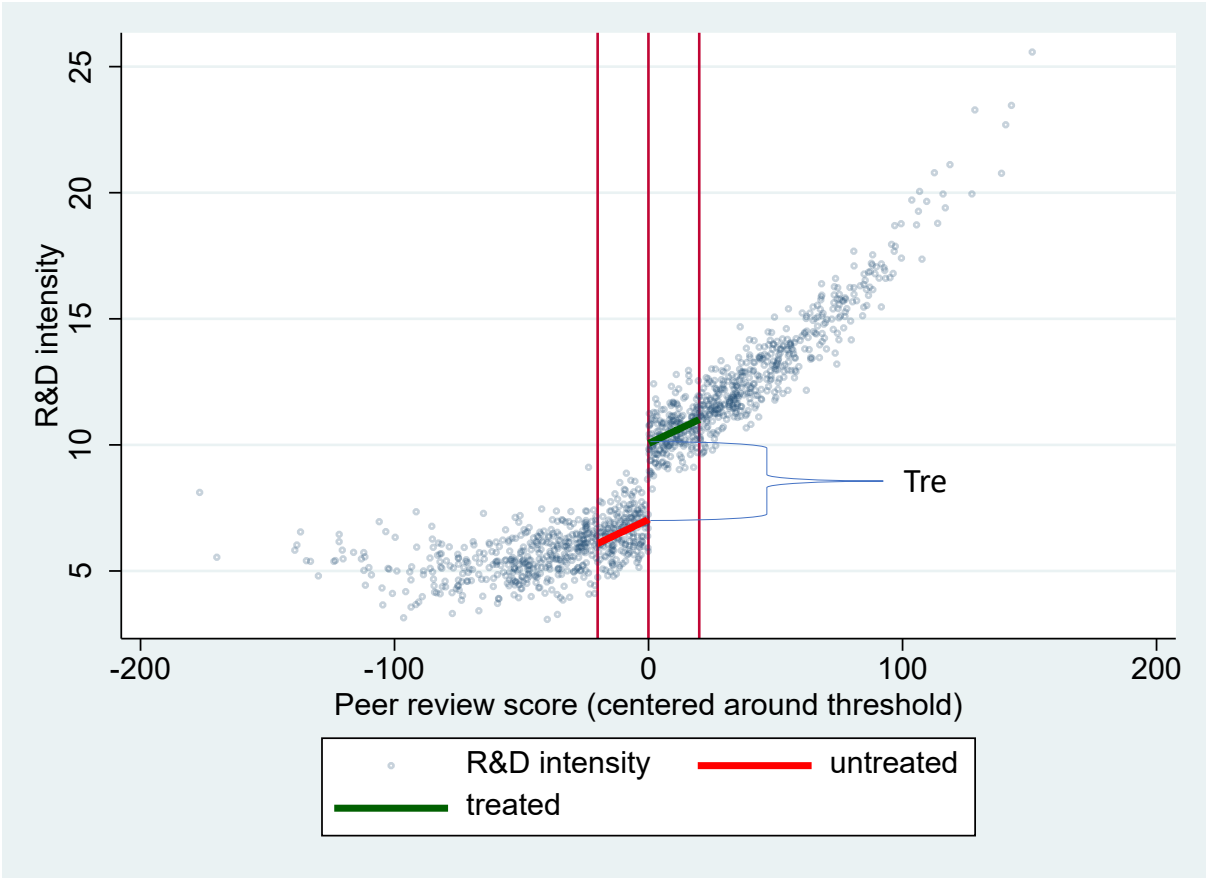
Taking into account multiple criteria that determine the treatment (e.g., combinations of age and size of a firm) is not easily possible.

2.3.4 Functional form

Another decision that has to be taken by the researcher is the choice of functional form on how the proposal scores enter the model, as the results of the RDD may be sensitive to this. For instance, one might choose a linear specification, and this may be appropriate closely around the threshold, but as the difference between the individual scores and the threshold increases, it can become imprecise.

See, for instance, Figure 8 below. In this example, regressions in the neighbourhood between -20 and 20 around the eligibility threshold have been performed. The correctly estimated treatment effect is the difference between the values of the two regression lines where they intersect with the threshold, the discontinuity. From eyeballing, however, one can see that the regression line for the untreated would be much more flat if all positive scores would have been used to fit the red regression line. The careful observer could find out that the red regression line could still have a correct fit across the whole range of R&D intensity for the funded firms if one would use the score and its squared value as independent variables. However, the correct choice of functional forms is often not that obvious. The same holds for the regression line of the treated observations. Even though less obvious, the correct functional form over the whole range of the data is a second order polynomial. Locally around the threshold, however, it can be approximated linearly – in this case as shown for scores between 0 and 20.

Figure 8: Hypothetical Example of RDD with a linear regression around the funding threshold



In principle, a non-parametric regression[13] could be applied, which makes no assumption on how the scores should be modelled. This, however, requires a very high number of observations which are often not available in the context of RDI programmes (but may well be in other fields such as labour market programmes for training).

The sample that is used for the RDD estimation is often restricted to a pre-defined neighbourhood closely around the threshold value In order to mitigate the functional form problem to a certain extent. Researchers usually try different cut-off values to restrict the sample and investigate the robustness of the results obtained because there is typically no clear definition of "closely" in this context. The idea behind this logic is that the

[13] The evaluator does not pre-specify a certain functional form for the relationship between the outcome variable and the explanatory variables in a 'non-parametric regression.' While this is a more general form of regression technique than, for example, the linear regression, the major disadvantage of non-parametric regressions is that very large amounts of data are required to obtain statistically sound results. These large data requirements are often not encountered in evaluations in the context of RDI policies.

more the sample is restricted around the threshold, the more functional form assumptions that are made are likely to hold *locally*, rather than for observations that are distant from the threshold value. The trade-off that is faced, then, is that an increasing chance that the model is correctly specified locally coincides with a loss in precision of the estimates as observations are discarded (i.e., the model might be more appropriately fitting the data, but the researcher may not find a statistically effect because of smaller sample sizes).

2.4 Instrumental Variable (IV) Regressions

Instrumental variable techniques are a very traditional approach to dealing with *endogeneity* problems that arise in regression models. For example, if one were to evaluate an R&D programme's effect, then one could specify an R&D investment model that yields an estimable equation that includes a programme participation variable. It is commonly believed that a linear regression, estimated with the OLS estimator, would lead to an overestimation of the programme effect because the group of participants is a selected sample that would have conducted more R&D than a randomly sampled control group, even in the absence of a policy. This creates an endogeneity problem of reverse causality because the policy outcome variable, R&D investment, may partially determine participation itself. Endogeneity problems like reverse causality result in inconsistent OLS estimates.

One solution to fix this problem include instrumental variable approaches. They are mostly implemented as Two-Stage-Least-Squares (2SLS) estimator. In this approach, the endogenous programme participation variable is first explained by all covariates, which also determine R&D investment, and additionally by one or more instrumental variables. The criteria for such instruments are demanding: (i) they have to be strongly and economically meaningfully correlated with the programme participation; and (ii) they must be exogenous to the investment model (i.e., they must neither depend on R&D investment themselves, nor should they have an effect on R&D investment independently of the programme participation).

The difficulty of fulfilling the two conditions above can be explained with an example: suppose the outcome variable of interest is R&D investment, and one has to find an instrumental variable for programme participation. One would need to think about what might determine programme participation. A variable that might come to mind is some measure of human capital among a firm's R&D staff. More qualified researchers might have better project ideas and might also be able to write better proposals. Thus, for example, the share of (R&D) personnel with a university degree could, among other variables, determine programme participation. However, it is obvious that such a variable will also determine R&D investment and, thus, also belongs in the regression of R&D investment on programme participation as a control variable. If personnel's qualifications are ignored, then the regression might mistakenly assign positive investment effects to the programme, instead of to human capital. Such thought experiments can be made for many variables and it often turns out that there are arguments for why a variable might also have to be included in the R&D investment equation or for why this variable might itself depend upon investment. Even when a possible candidate is found that complies with condition (ii) mentioned above, it might often fail to satisfy condition (i) mentioned above. If a variable is not highly correlated with programme participation, then it is a so-called *weak instrument*. Using weak instruments in a 2SLS regression can lead to substantial bias in the estimates, thereby rendering them unreliable. Whether a variable fulfils the condition of strong correlation can be tested empirically, but it cannot be known a-priori.

Identifying variables that can serve as candidates for instrumental variables is often a practically challenging problem in the context of RDI programmes. Therefore, IV estimators are not commonly used in RDI evaluation studies even though they are very commonly used in economic research in general. One exception is policy evaluation studies in which special programme features can be exploited to form instrumental variables, such as unanticipated eligibility changes over time or participation thresholds that are out of the firms' control (see the previous subsection on RDD).

2.5 Randomised Control Trials (RCT)

RCTs circumvent the problem of endogeneity by creating an experimental setting in which the treatment (i.e, the R&D subsidy) is randomly allocated to firms, thereby eliminating systematic differences among treated and control observations. RCTs are best known from their application to clinical trials. Even though RCTs may often be considered as the gold standard in the field of evaluation studies, serious ethical concern have been raised about many applications. While this might be more obvious in the context of clinical trials that employ human subjects, ethical concerns are also present in RDI programmes. The treatments may lead to a significant

distortion of the current market structure. For example, a firm might develop a product with the public grant that would have never received support if a peer review process would have been applied because it "only" increases private profits, but the expected social returns are low. Instead, it might kill jobs in other firms on the market (i.e., the negative social returns might outweigh the positive ones). In addition, a random allocation of innovation subsidies may be considered as inefficient wasting of public resources because R&D grants are typically non-negligible in their amounts and easily range from between EUR 100,000 and EUR 250,000 in standard programmes, even for projects conducted by small firms.

3 The Evaluation of R&D tax credit schemes

This guidebook has thus far used R&D grant schemes as examples for CIE studies in the area of public RDI support programmes. The methods discussed can also be used for the evaluation of R&D tax credit schemes, but with one limiting factor: the credible construction of a control group can be more challenging in the context of R&D tax credit schemes because all R&D performers are typically eligible for tax deductions when such a scheme is in place. Thus, there is usually no control group of firms available that participated in the policy.

To that end, scholars often use within-scheme changes that appear over time, instead of using non-participants. For example, Belgium introduced a scheme based on an exemption from advance payments surrounding the withholding tax of R&D personnel's wages. The scheme was initiated in 2006 and only the wages of R&D personnel, who held a Ph.D. degree, were eligible for the deduction of advance tax payments. In 2007, the scheme was extended to R&D employees holding master's degrees. Thus, the policy change between 2006 and 2007 can be interpreted as a treatment that benefits those firms that had R&D personnel with master's degrees. Firms that only had other R&D personnel could serve as a potential control group.

Similarly, Guceri and Liu (2019) investigate the effects of the UK R&D tax credit programme. This programme applies different deduction rates for SMEs and large firms. In 2008, the UK government changed the eligibility conditions such that medium-sized firms qualified for more generous rates. Until 2008, SMEs could tax deduct £150 for every £100 spent on qualifying R&D. Large companies could deduct £125 for every £100 of R&D. In 2008, these deductions changed to £175 for SMEs and £130 for large firms. Furthermore, an SME was formerly a firm that had no more than 250 employees according to the scheme's definition. From 2008, this threshold was increased from 250 to 500. Thus, the medium-sized firms benefit relatively more from the scheme than larger firms. Formerly, a medium-sized firm (i.e., 250-500 employees) could tax-deduct £125, but this was raised to £175. This implies that the cost of conducting R&D was reduced much more relatively than for large firms.

Guceri and Liu use this discontinuity in the programme to estimate DiD models on R&D investment where the medium-sized firms are a treatment group and the larger firms are a control group. Their data comprises all UK R&D performers between 2002-2011. They find that the medium-sized firms, the treatment group, increased their R&D spending by 33% in response to the policy reform. Transforming their estimates into a "bang-for-the-buck" interpretation, by means of a back-of-the-envelope calculation, yields that for each £1 forgone in tax revenues, £1 to £1.5 additional private R&D spending was generated. This implies that the policy is very effective and crowding-out effects do not seem to be present.[14]

[14] For an alternative evaluation approach to the UK R&D tax credit programme using production functions, see Devnani et al. (2019).

4 How to set up a counterfactual impact evaluation (CIE)

Setting up a counterfactual impact evaluation study requires some careful balancing between desired outcomes (i.e., the evaluation questions that ideally should be addressed and the reality of data availability). While the econometric evaluation estimators' levels of sophistication may suggest that the questions that can also be answered by using similarly sophisticated econometric techniques, the opposite is the case. Applying such econometric estimation techniques successfully requires a large amount of (high-quality) data on programme beneficiaries and, in most cases, also on firms that did not obtain an award or did not even intend to participate in the programme. It is advisable to limit the evaluation questions to a small number of core points because what can be measured by variables that are widely available for firms in the target population of interest is typically rather limited. The following subsections provide examples of commonly used data sources and mention some typical questions that are raised in CIEs.

4.1 Data requirements

4.1.1 Programme participants

In order to conduct a CIE for an R&D, innovation grant scheme, or similar programme, it is required that information on the programme participants be made available. At the very minimum, the following information should be made available by the public agency managing the programme at the project level (i.e., per application for the grant programme):

- Full company name incl. legal form;
- Company location (full address or at least the city);
- Company ID: typically a VAT ID of the firm or other commonly used unique identifier;
- Date of application for the grant;
- Dates of project beginning and ending;
- Total cost of the submitted project (if multiple applicants, per applicant);
- Subsidized amount or rate of total cost (if multiple applicants, per applicant);
- If multiple programmes or subprograms, the specific subprogramme or field;
- Information on whether the application was subject to any bonus or preferred treatment (e.g., micro and small firms or applications with universities in the consortium sometimes get a higher subsidy rate);
- If available, scores from a peer review.

Ideally, the managing agency would also systematically collect the information from applications that were not successful (i.e., where no subsidy was granted).

4.1.2 Control group and further firm-level data

In addition to the essential programme participation information, it is required to collect data on possible outcome variables that should be subject to the evaluation, as well as about other control variables that may influence the chosen outcome variables, independently of the programme participation. Which variables are relevant to collect will, hence, depend on the type of programme and on the evaluating agency's specific interest.

It is important to note that such data have to be collected for all programme participants and for non-participants that could later form a control group in the empirical CIE study.

For instance, for the most standard evaluation question on whether a policy scheme increased R&D or innovation inputs in the beneficiary firms, one would typically collect information on:

- possible outcome variables, such as:
    - R&D expenditure (possibly split by internal and external R&D);
    - Total innovation expenditure;

    - R&D employment.
- and possible control variables, such as:
    - firm size (employment, total assets, or sales);
    - economic sector;
    - capital intensity;
    - internal financial resources;
    - debt levels;
    - information on corporate governance (e.g., stand-alone firm, subsidiary of a group, parent company);
    - other variables that partly determine innovation inputs (e.g., regional characteristics).

If innovation output variables are to be considered as outcome variables, then standard examples are:

- indicators on whether a firm introduced a new product or service;
- indicators on process innovation;
- the sales achieved with new products (or market novelties);
- new patent applications that the firm filed;
- less commonly used but, in principle, trademarks and registered designs (such variables could be used to approximate measures of new products and services).

It is not uncommon to also use more general firm performance indicators as outcome variables. Examples include:

- total sales (growth) or value-added (growth);
- employment (growth);
- (change in) market value (measuring market values is only possible for publicly traded firms that – although large important companies – are a minority in continental Europe in terms of their share in total numbers of firms in a country).

Box 4. Control group

Generally, firms in the control group should have a similar likelihood to participate in the respective programme as the awardees.

It seems ideal, therefore, to use applicants that were not successful in obtaining a grant.

While this is theoretically a sound and attractive approach in most cases, it often turns out in practice that not very many proposals get rejected in the end in national or regional RDI programmes. It is not uncommon that 80%-85% of applications are funded and only 15%-20% are rejected. In such cases, relying only on rejected applicants results might possibly yield samples that are too small for a CIE. However, unsuccessful applicants as a control group can be supplemented with non-applicants that were eligible to apply.

4.1.3 Typical data sources

Where might we obtain data for a CIE? While the programme participation data should be supplied by the managing agency, it is usually not the case that other firm-level data are systematically held in databases by the managing authorities, especially for non-applicants.

4.1.3.1 Community Innovation Surveys

The Community Innovation Survey (CIS) is the European reference survey on innovation in enterprises. The EU countries first introduced the survey in 1992 and, since then, it has become a regular biennial data collection.[15] At present, the survey is carried out in the EU, EFTA and in the candidate countries. However, the CIS's format was also adopted in several other countries, among others Canada, Chile, Brazil and South Africa.

The guidelines for data collection are laid out in the so-called Oslo Manual (OECD/Eurostat, 2018). The firm-level data in the CIS are usually stratified random samples that are representative of the target population of firms in the country.[16] The CIS does not cover all sectors of the economy, only those ones that are typically relevant for CIEs of RDI programmes (i.e., the manufacturing and business service sectors).

The surveys include innovation input and output variables, as mentioned above. The CIS are often the only source for such variables and is, therefore, often used in CIEs.

4.1.3.2 R&D Surveys

Like CIS surveys, the European Member States and other countries also conduct so-called "R&D surveys" to collect detailed information about research and development efforts in the business sector. The results of these surveys provide important parts of the official OECD R&D statistics.[17]

In addition, these surveys are also a rich data source for CIEs. The Frascati Manual (OECD, 2015) constitutes the international guidelines on how to collect R&D data, and it contains descriptions of the core variables that are available in the data. The advantages and disadvantages of the data are very similar to the CIS and, therefore, are not listed separately here.

4.1.3.3 Orbis global database

A frequently used database for CIEs in the R&D context is the Orbis global database[18] that is maintained by Bureau van Dijk (https://www.bvdinfo.com/en-gb/). The Orbis database contains firm-level information such as balance sheet data and profit-and-loss sheet data from a large set of countries in the world. It is commonly used as source of information on variables such as sales, value added, employment, assets, cash-flow, equity, debt and many more variables besides. It also contains, among others, information about the ownership of companies. Even though it does not contain R&D and innovation data, it is often used to add variables to a CIE in which other CIS data are used. It is often the only source of data at the firm level because it is widely accessible, unlike the CIS and R&D surveys. In such cases, scholars have to use general firm performance variables as potential outcome variables of the RDI policy scheme, such as sales or value-added (growth), or jobs created (measured as total employment change).

In some cases, researchers may have access to (even more detailed) databases at the national level which are often the input for the Orbis global database.

---

[15] The CIS data are also available for researchers at Eurostat, but only in anonymized form: https://ec.europa.eu/eurostat/web/microdata/community-innovation-survey.

[16] One exception is Italy where the survey is sent to the whole target population of firms and where responding is mandatory. This has led to a response rate of roughly about 80% in the past (even though the survey is mandatory, firms may still not respond and are prepared to pay a fine instead).

[17] Cf. https://www.oecd.org/sti/inno/researchanddevelopmentstatisticsrds.htm

[18] A former European version of the database was called "Amadeus". However, BvD stopped offering this product and integrated the European data into the Orbis global database.

Box 5. Advantages and disadvantages of the CIS

Advantages:

Includes many variables relevant for CIEs of RDI programmes;

Often the only source for relevant information on innovation;

Also includes firms that can be used as control group.

Disadvantages:

The CIS questionnaires are sent to a sample of firms in the target population (i.e., one will never find all programme participants in the CIS, but instead only a (small) fraction). Some countries anticipate that CIEs will be conducted and, therefore, add known programme participants to the random samples that receive the survey questionnaires,

The European-wide CIS data are available at Eurostat, but only in anonymised form. This database can, therefore, not be used for CIE as programme participants cannot be identified ex-post,

Researchers often have to approach the CIS-implementing institution in their country for conducting CIEs, because when collaborating with the CIS-implementing institution directly, the implementing institution can identify programme participants as they have the non-anonymised data. However, the data access conditions and procedures for merging information such as programme participation with the CIS data are often cumbersome for the researchers outside of the CIS-implementing institution.

4.1.3.4 PATSTAT database

The PATSTAT database contains the population of patent applications from 90 patent issuing authorities including the most important national patent offices and the European Patent Office. There are over 100 million patent applications recorded in the database at present.[19]

Patent data are frequently used as outcome variable to measure inventive activity or inventive success at the firm-level; past patent application (i.e., the patent stock of a firm) are often used as a measure for past (successful) engagement in R&D activities.

While not all inventions are patented, patents are a very useful data source for CIEs because the patent documents leave a paper-trail that can be quantified, and it is widely available for all firms in most industrialised economies.

4.1.3.5 Trademark data

Much less researched than patent data are trademark data. Trademarks are also widely available for Europe through the European Union's Intellectual Property Office's OHIM database.[20]

Trademarks are signs that allow customers to identify products. Such signs may be, among others, words, figures, shapes, patterns and sounds.[21] Thus, newly registered trademarks may also be used to approximate product companies' innovations and, therefore, variables constructed out of trademark registrations may supplement patent data for the measurement of innovative activity in the business sector. This is particularly so for sectors in which patenting does not play an important role, such as business services; trademark data could also be interesting sources of information for CIEs.

4.1.4 Secondary data sources versus primary data collection

It seems appropriate to conduct an own primary data collection through a survey for desired CIEs because secondary data sources for specific evaluation questions are frequently lacking or are difficult to access (at least in a non-anonymised form).

[19] Cf. https://www.epo.org/searching-for-patents/business/patstat.html
[20] Cf. https://euipo.europa.eu/ohimportal/en/databases
[21] Cf. https://euipo.europa.eu/ohimportal/en/web/quest/trade-mark-definition

While this is certainly possible, it should be noted that a survey is highly costly to conduct. While a written survey is possibly more expensive than either a Computer-Aided Telephone Survey (CATI) or an online survey, implementing either is far from trivial. In any case, a questionnaire must first be designed carefully. The questions to be asked should usually be pre-tested in face-to-face interviews with a selected number of firms in order to ensure that the respondents understand the questions as intended and that the variables constructed from the answers do not mismeasure the underlying economic activity. While a "paper-and-pencil" survey entails significant printing cost for the questionnaire, the shipping cost of the postage including the return envelopes, CATIs and online surveys require specialised software solutions that may also be costly when only purchased for one specific CIE. In addition, a very practical and often time-consuming challenge is to collect relevant e-mail addresses from contact people at firms who could fill in the questionnaire. If e-mails are sent to general e-mail addresses that are not directly linked to a person, then the response will usually be poor.

Conducting surveys also takes a significant amount of time. It is not uncommon that a survey will be in the field for multiple months. After the questionnaires have been tested, and the first questionnaires have been sent, some time has to be allowed for the respondents to answer the survey. After some weeks, one typically takes stock of the responses and sends out reminders for which one has to allow response time once again. It is common to send out two reminders.

Furthermore, surveys entail the risk that the response rate will be very low. It may turn out that the survey data are not rich enough to conduct a rigorous CIE in the end. Therefore, smaller surveys are often used only as a supplement to secondary data sources for more qualitative evaluations.

4.2 Setting up a CIE

It is advisable to follow a few rules when designing the policy scheme and when preparing (the call for tenders for) the study in order to successfully set-up an econometric CIE. The policy design is most important for the success of a policy scheme and a credible evaluation, as is the selection of independent evaluators who are free of possible conflict of interests. Ideally, a detailed evaluation plan will already be prepared before a scheme is launched. In such an evaluation plan, possible evaluation questions about whether and how the policy scheme might meet its objectives should be formulated. Furthermore, possible variables and ideas on how to collect data for their measurement will usually be included in such a plan.

4.2.1 Policy design

4.2.1.1 Intervention logic and meaningful outcome variables

It is important to think about the intervention logic when designing a policy scheme for RDI. The most standard intervention logic is driven by the market failure and financial constraints arguments described previously: public agencies support activities in firms in order to reduce the firms' cost of R&D and innovation because it is commonly believed that R&D efforts in the private sector are below the socially desirable level.

One might expect a positive effect on input factors such as R&D expenditure, R&D employment or total innovation expenditure because firms should use public RDI grants to increase their efforts. If no statistically significant positive effects on inputs can be found, then possible crowding-out effects within the policy scheme cannot be rejected.

The effects that should be expected might be much less clear for other choices of outcome variables. Bronzini and Iachini (2014) considered assets as outcome variable (see the example of an RDD study in section 2). Given the information in Table 1 of this guidebook, it seems questionable whether assets are an appropriate target variable as an outcome of an RDI support programme. If most firms' R&D expenditure flow into wages of R&D personnel and current cost of R&D, then why would it be informative for the efficacy of a programme to investigate assets? The expected effect, if any, can only be very small because investment into capital goods, such as equipment, is a rather small share of total R&D expenditure.

Furthermore, what to expect regarding *output additionality* might also be ambiguous, a concept that is often considered in evaluation studies. For example, scholars analyse whether the additional R&D inputs also lead to greater innovation output. Here it is already critical to define the hypothesis to be tested very precisely because establishing such a relationship is far from trivial. Among other complexities that have to be dealt with, the following example may highlight the view that careful hypothesis development is of utmost importance in CIEs. In *input additionality* studies; it is popular to summarise the results as "bang for your buck", that is: how many *privately financed* additional EUR in R&D investment are triggered by 1 EUR of *public R&D support*; a value large

larger than 0 but smaller than 1 (not full crowding-out but partial crowding out) or a value larger than 1 EUR (no crowding-out)?

Such a logic is often also applied to *output additionality*. For example, how many additional sales with innovative products per EUR of R&D are achieved with public support? If one follows neoclassical economic theory, then it is commonly believed that factor inputs have decreasing marginal returns (i.e., every additional EUR invested in R&D will have less returns than the previously invested EUR). In this logic, it would be wrong to expect that the ratio of sales with new products to R&D in publicly supported firms should exceed the comparison ratio defined by the counterfactual situation of the policy being absent. Instead, it may be expected that this ratio falls as the additional R&D triggered by the subsidy might have lower marginal returns to some extent. Therefore, a more meaningful hypothesis test might be to specify that the productivity of R&D in subsidised firms should not be significantly lower when compared to the counterfactual situation of not having participated in the programme.

The problematic evaluation of output additionalities can be seen in a recent evaluation of the Finnish R&D programme run by Tekes (see Fornaro et al., 2020). The authors employ a conditional-DiD methodology to a panel of Finnish firms and find a strong effect for input additionality in terms of R&D spending and R&D employment. When turning to output additionality, where labour productivity is chosen as the outcome variable, the analysis remains inconclusive. The value-added per employee remains quite constant over time (i.e., there is no significant increase in the post-treatment phase).

4.2.1.2 Eligibility rules

It might be useful to think about eligibility rules when initially designing the policy scheme. In many cases, it is not useful that all firms in the business sector can apply to the programme or that all firms are entitled to receive the same benefits.

For instance, policy schemes that are mainly motivated by financial constraints arguments, such as Eureka's Eurostars programme or the Italian Start Up Act, typically impose thresholds on firm size, age and also R&D intensity in the latter example. In other R&D schemes, the subsidy rate varies with firm characteristics (e.g., firm size), or with attributes of projects (e.g., for example, basic versus applied research or collaboration with scientific institutions).

Such eligibility criteria may be used to define thresholds that can separate the treatment group exogenously from a possible control group that could be used for a CIE. Scholars often believe that a change in eligibility criteria, induced by the agency as an exogenous shift in eligibility criteria, may well be used as identifying criterion for candidates of a control group (e.g., firms that were not eligible prior to the change in participation rules, but which then become eligible and who might apply a good control group for the time period before the change).

4.2.1.3 Data collection

It is also useful to think about the options for systematic data collection when designing the policy scheme and application procedure. One might even require firms to report desirable information for a CIE with the application (e.g., R&D conducted in previous periods).

A monitoring system for the duration of the projects and a post-participation period could also be implemented. While certainly useful, detailed data collection from the participants has limitations in the context of CIEs, as: (a) these numbers are obviously self-reported in the context of an application and, therefore, might not be fully trustworthy; and (b) scholars often need corresponding information from a non-supported control group of firms.

In the ideal case, possible applicants and a control group would be embedded in regular data collection efforts that are not directly linked to the programme participation, such as the Community Innovation Surveys and the R&D surveys, or similar data collections that are not available in administrative sources, such as accounting data or patent data, are desired.

4.2.2 Preparing a (call for tenders for a) CIE

It is advisable to consider the following view when preparing a CIE:

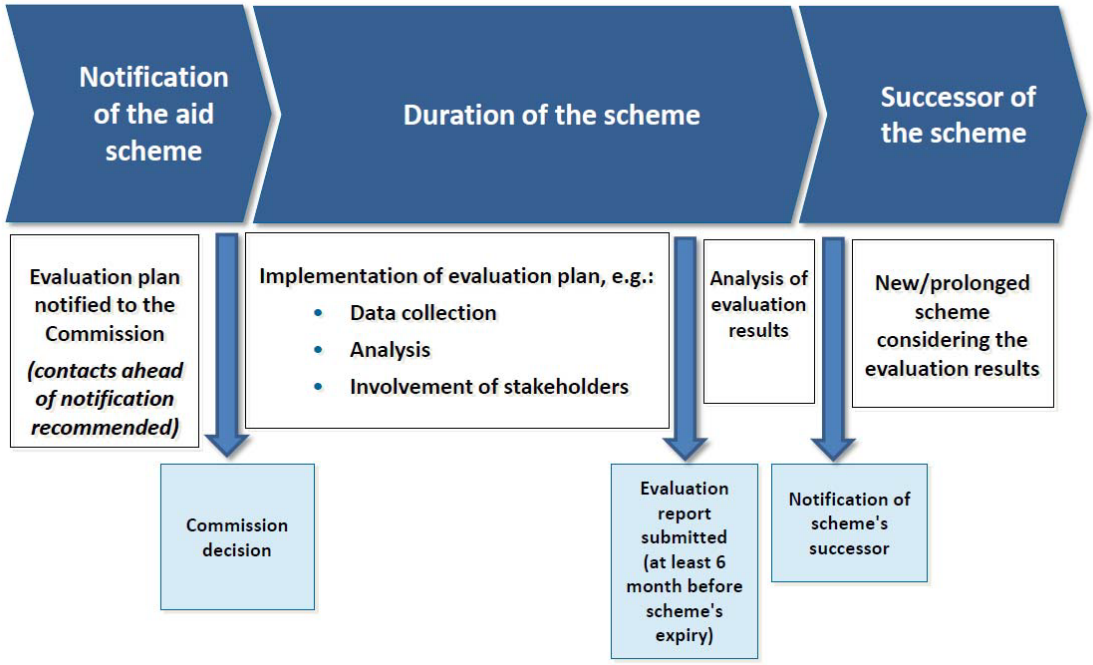
- A typical European policy evaluation study asks over a dozen evaluation questions and involves both qualitative and quantitative parts, and possibly even an econometric CIE among the quantitative part

The advantage of an econometric CIE is its methodological rigour that is used to establish *causal effects*. The price for the methodological rigour is a high level of technical sophistication in terms of statistical and, in particular, econometric methods and the need for detailed data at a large scale. Therefore, the evaluation questions to be addressed with a CIE can typically only have a very basic and straightforward nature (e.g., "how many R&D employees would the supported companies have had on average if they were not supported?") in order to derive a program's causal effect, such as the treatment effect on the treated.
The insights of the CIE usually have to be combined with more qualitative evidence, such as interviews with participants, programme managers and other stakeholders to form qualified conclusions about broader, less stringent evaluation questions;

- A reasonable timing of the CIE is important. It is often the case that legal frameworks stipulate an evaluation during or directly after a certain funding round within a policy scheme. Hard economic data at the firm level are necessary for a CIE in the context of RDI schemes, however, including information on innovation inputs and outputs and general firm performance variables. These two matters often create difficult problems. In order to successfully apply a CIE, as many programme participants as possible should be taken into account and some time should elapse in order to measure their innovation inputs and outputs. For example, an innovation survey (e.g., the CIS) that was carried out in the year 2020 includes data for inputs and outputs for the year 2019 (and some for 2018). However, it usually takes until the year 2021 before the cleaned survey data are published and become available for evaluation studies. This means that a programme could be evaluated concerning innovation inputs in 2019 at the earliest in the year 2021. If, however, innovation outputs were to be a part of the CIE, then one might want to wait until more time has elapsed for inputs to be transformed into outputs. Patent data are a common example: legally, a patent application is published after 18 months after its submission to the patent office. However, in addition, there is an input lag of the application document into the patent databases so that three years can easily pass before the patent application databases are complete for any given year. If further information is required, for instance whether a patent application has been granted, then another three to five years might need to be added in order to observe the majority of patent applications outcomes.

The timing dilemma becomes clear when comparing the current State aid rules to actual data availability in practice. When designing and implementing a scheme, EU Members States have to notify the EC about the intended scheme and must present an evaluation plan. The scheme is supposed to be implemented only after the evaluation plan has been approved and an evaluation is supposed to be conducted during the duration of the scheme. An evaluation report should be submitted at least 6 months prior to the expiration of the scheme. Figure 9 shows the evaluation timeline, as suggested by the European Commission.

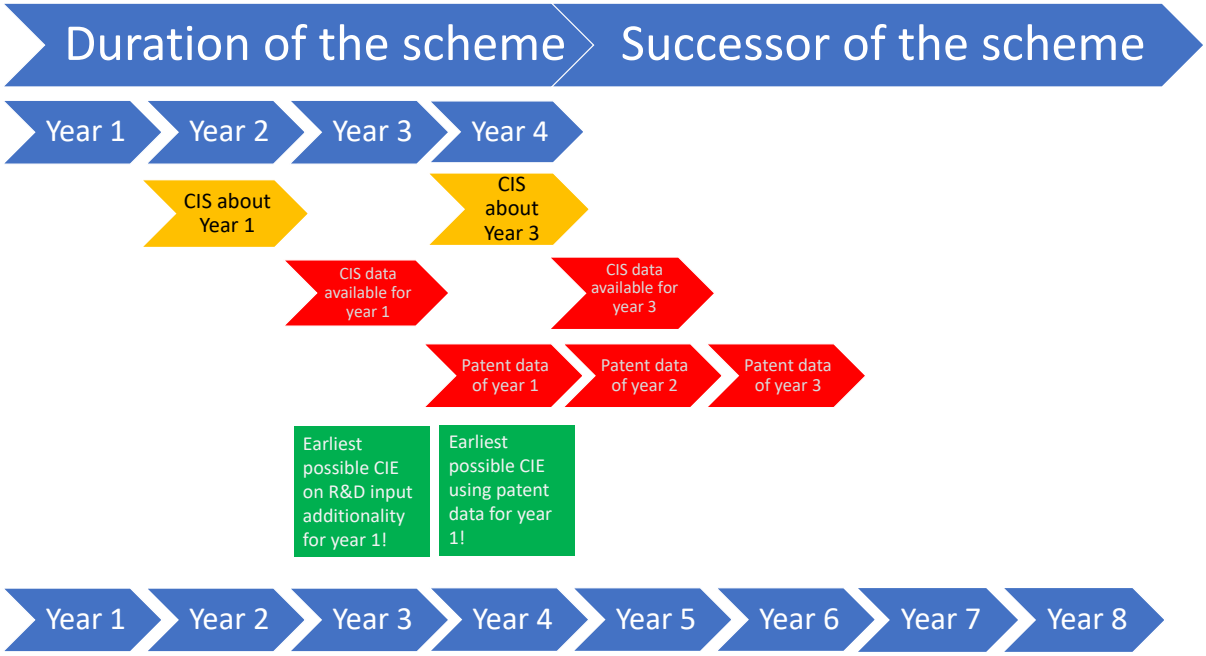
Figure 9: Suggested evaluation timeline by the European Commission



Source: European Commission (2014).

Unfortunately, this proposed timeline often stands in contrast to data availability for such evaluations. Figure 10 sketches the timeline for a hypothetical programme. Suppose the evaluator would like to use data from the Community Innovation Surveys (CIS) and possibly even patent data. In each year that the CIS is carried out, it collects information on the previous year, and in most countries the CIS is only carried out every second year. The data then typically become available for further usage in the subsequent year. This implies that a policy scheme running for four years can, at the earliest, be evaluated in year three of its duration. Then, only the first year of the programme can be evaluated using CIS data. If a further year of the programme were to be included in the CIE, then the study could only be conducted in the fifth year after the programme started. While patent data have the advantage that they are available for every year, the earliest year when any data from the programme period become available is the fourth year.

Figure 10: Example of timeline of scheme and data availability



Such timelines imply that a programme cannot be evaluated comprehensively within the duration thereof. Some years have to pass until rich enough data is available for CIEs evaluating the whole programme. At least two full years have to elapse until even the first year of the programme can be investigated regarding input additionally.

5 Conclusion

This guidebook has attempted to explain the need and virtues of econometric counterfactual impact evaluations (CIE) of State aid schemes for research, development and innovation as well as the main intuitions behind the most commonly used methods, their data requirements and limitations.

The research on methods for CIE is progressing quickly and the methods have become increasingly and strive to establish causality between a policy intervention and firm-level outcomes more convincingly. Moreover, data requirements become more demanding and, therefore, the applicability in everyday policy practice becomes more challenging. It is a difficult task for the CIE's managing agency to balance the claim of causality and the feasibility of the CIE. This guidebook aims to deliver some guidance on how to achieve the goal of including meaningful CIE in broader policy evaluation studies.

References

Almus, M., and Czarnitzki, D. (2003). The effects of public R&D subsidies on firms' innovation activities: the case of Eastern Germany. Journal of Business & Economic Statistics, 21(2), 226-236.

Arrow, K.J., (1962). Economic welfare and the allocation of resources for invention. In: Nelson, R.R. (Ed.). *The Rate and Direction of Inventive Activity: Economic and Social Factors*, National Bureau of Economic Research, Conference Series. Princeton University Press, Princeton, pp. 609–625.

Ashenfelter, O. (1978). Estimating the Effect of Training Programs on Earnings. *Review of Economics and Statistics*, 60, 47-50.

Biancalani, F., D. Czarnitzki and M. Riccaboni (2022), The Italian Startup Act: A Microeconometric Program Evaluation, *Small Business Economics*, 58, 1699–1720.

Bloom, N., Van Reenen, J., & Williams, H. (2019). A toolkit of policies to promote innovation. *Journal of economic perspectives*, *33*(3), 163-84.

Bronzini, R., and Iachini, E. (2014). Are incentives for R&D effective? Evidence from a regression discontinuity approach. *American economic journal: economic policy*, *6*(4), 100-134.

Conti, M., Ferrara, A., Ferraresi, M., and Giua, L. (2021), Regional state aid: a toolbox on counterfactual impact evaluation, JRC Technical Report, European Commission, Ispra, 2021, JRC125911.

Crivellaro E., Ferrara, A., Ferraresi, M., and Giua, L. (2021), Support to the quality assessment of evaluation plans and reports in the area of State Aid (EVALSA II): the 2021 activity report, JRC Technical Report, Ispra: European Commission.

Czarnitzki, D. and H. Hottenrott (2011a), Financial Constraints: Routine versus Cutting Edge R&D Investment, *Journal of Economics and Management Strategy* 20(1), 121-157.

Czarnitzki, D. and H. Hottenrott (2011b), R&D Investment and Financing Constraints of Small and Medium-Sized Firms, *Small Business Economics* 36(1), 65–83.

Czarnitzki, D., Hünermund, P., & Moshgbar, N. (2020). Public procurement of innovation: evidence from a German legislative reform. *International Journal of Industrial Organization*, *71*, 102620.

Czarnitzki, D., and Lopes-Bento, C. (2013). Value for money? New microeconometric evidence on public R&D grants in Flanders. *Research Policy*, *42*(1), 76-89.

Devnani, S., R. Ladher and N. Robin (2019), Evaluation of the Research and Development Tax Relief for Small and Medium-sized Enterprises, London.

European Commission (2014), Common methodology for State aid evaluation, Commission Staff Working Document SWD(2014) 179 final, Brussels.

Farrel N., Energy State aid: A Toolbox on Counterfactual Impact Evaluation, eds: Crivellaro E., Ferrara A., Ferraresi M., Giua L., Granato S., Lapatinas A., Vidoni D, EUR 31105 EN, Publications Office of the European Union, Luxembourg, 2022, ISBN 978-92-76-53298-9, doi: 10.2760/5715, JRC129621.

Fornaro, P., H. Koski, M. Pajarinen and I. Ylhäinen (2020), Evaluation of TEKES R&D Funding for the European Commission – Impact Study, Helsinki: Business Finland.

Guceri, I., and Liu, L. (2019). Effectiveness of fiscal incentives for R&D: Quasi-experimental evidence. *American Economic Journal: Economic Policy*, *11*(1), 266-91.

Hall, B.H. (2002), The Financing of Research and Development, *Oxford Review of Economic Policy* 18(1), 35-51.

Hall, B.H. (2008), The financing of Innovation, in: S. Shane (Ed.), *Handbook of technology and innovation management*, Oxford: Blackwell, 409-430.

Hall, B.H. and J. Lerner (2010), The Financing of R&D and Innovation, in: B.H. Hall and N. Rosenberg (Eds.), Handbook of the Economics of Innovation, North-Holland: Elsevier, 609-639.

Himmelberg, C. P., and B.C. Petersen (1994), R&D and internal finance: A panel study of small firms in high-tech industries, *Review of Economics and Statistics* 38-51.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American statistical Association*, *81*(396), 945-960.

Hovdan, B., Czarnitzki; D., Pierluigi, A. (2023) R&D tax credits: A review and an econometric study on the policy mix. In: Braunerhjelm, Andersson, Blind, Eklund (eds.) *Handbook of Innovation and Regulation*, Edward Elgar Publishing.

Hünermund, P., & Czarnitzki, D. (2019). Estimating the causal effect of R&D subsidies in a pan-European program. *Research Policy*, *48*(1), 115-124.

Imbens, G. W., & Wooldridge, J. M. (2009). Recent developments in the econometrics of program evaluation. *Journal of economic literature*, *47*(1), 5-86.

Kaufmann, P. , B. Bittschi, H. Depner, I. Fischl, J. Kaufmann, E. Nindl, S. Ruhland, R. Sellner, V. Struß, T. Vollborth, J. Wolff von der Sahl (2019), Evaluation des Zentralen Innovationsprogramms Mittelstand (ZIM) Richtlinie 2015 Endbericht, Studie im Auftrag des Bundesministeriums für Wirtschaft und Energie (Berlin), Wien.

Mansfield, E., John, R., Anthony, R., Samuel, W., George, B., (1977), Social and private rates of return from industrial innovations. *Quarterly Journal of Economics* 91 (2), 221-240.

Nelson, R.R., 1959. The simple economics of basic scientific research. *Journal of Political Economy* 49, 297–306.

NOVADAYS – UCM (2020), Impact evaluation study of the aid scheme on CDTI R&D projects – Report commissioned by the Centre for the Development of Industrial Technology, Madrid.

OECD (2015), Frascati Manual 2015: Guidelines for Collecting and Reporting Data on Research and Experimental Development, The Measurement of Scientific, Technological and Innovation Activities, OECD Publishing, Paris.

OECD/Eurostat (2018), Oslo Manual 2018: Guidelines for Collecting, Reporting and Using Data on Innovation, 4th Edition, The Measurement of Scientific, Technological and Innovation Activities, OECD Publishing, Paris/Eurostat, Luxembourg, https://doi.org/10.1787/9789264304604-en

Smith, J. A., & Todd, P. E. (2005). Does matching overcome LaLonde's critique of nonexperimental estimators?. *Journal of econometrics*, *125*(1-2), 305-353.

Szücs, F. (2018). Research subsidies, industry–university cooperation and innovation. *Research Policy*, *47*(7), 1256-1266.

Zúñiga [illegible] (2012). Assessing the effect of public subsidies on firm R&D investment: a survey. *Journal of economic surveys*, *28*(1), 36-67.

List of abbreviations and definitions

Control group	Counterfactual analysis requires finding the most comparable control group (i.e., a group of firms which should be as similar as possible to the group of units that received the aid — except that they have not benefitted from that aid).
Counterfactual	To estimate the effect of the aid on the aid beneficiaries, it is necessary to construct a 'counterfactual' (i.e., to establish a reasonable scenario capturing what would have likely happened to the aid beneficiaries if they had not received it).
Covariate	Explanatory variable or independent variable. It is a variable (on the right-hand side of a regression equation) that allows explaining the dependent variable (Y) on the left-hand side of the estimated equation.
Difference-in-Differences (DiD)	DiD compares the performance of a treated sample pre- and post-treatment relative to the performance of a control group pre-and post-treatment.
Evaluation	The systematic collection and analysis of information about programmes and projects, their purpose and delivery; it derives knowledge on their impact as a basis for judgments. Evaluations are used to improve effectiveness and inform decisions about current and future programming.
Impact	The change that can be credibly attributed to an intervention. Same as 'effect' of intervention or 'contribution to change'.
Input	An effort that a firm undertakes to achieve a certain goal (e.g., R&D investment is an input factor in the overall innovation process).
Intervention logic	Describes the needs and problems the scheme intends to address, the target beneficiaries and investments, its general and specific objectives and the expected impact.
Instrumental Variable (IV):	An instrumental variable is a third variable introduced into regression analysis that is correlated with the predictor variable, but uncorrelated with the outcome variable. By using this variable, it becomes possible to estimate the true causal effect that some explanatory variable has on an outcome variable.
Market failure	A market failure is a situation in which the allocation of goods and services by a free market is sub-optimal, often leading to a net loss of welfare.
Propensity Score Matching (PSM)	The propensity score allows one to design and analyse a non-randomised study so that it mimics some of the particular characteristics of a randomised controlled trial. The propensity score is a balancing score. Conditional on the propensity score, the distribution of observed baseline covariates will be similar between treated and untreated firms.
Outcome variable	This refers to the target variable of a policy. Outcome variables could be both input and outputs.
Output	The result of a production process (e.g., innovation outputs could be sales with new products that have been developed and introduced to the market recently).
Randomised controlled trial (RCT)	A randomised controlled trial is a form of scientific experiment in which participants are randomly allocated among compared treatments. Provided it is designed well, conducted properly, and enrols enough participants, an RCT can isolate the causal effect of a treatment (e.g., an R&D grant) on an outcome (e.g., a firm's R&D investment).
Regression	A statistical technique that relates a dependent variable to one or more (independent) explanatory variables. A regression model is able to show whether changes observed in the dependent variable are associated with changes in one or more of the explanatory variables.
Treated group	Treatment group or intervention group or programme participants. It is the group of firms that receives the intervention. This is then compared to the comparison or control group, which, on the contrary, does not receive the treatment.
Treatment	Intervention. It is the project, program, aid or, more in general, the policy to be evaluated.

List of boxes

List of figures

List of tables

GETTING IN TOUCH WITH THE EU

In person

All over the European Union there are hundreds of Europe Direct centres. You can find the address of the centre nearest you online (european-union.europa.eu/contact-eu/meet-us_en).

On the phone or in writing

Europe Direct is a service that answers your questions about the European Union. You can contact this service:

— by freephone: 00 800 6 7 8 9 10 11 (certain operators may charge for these calls),

— at the following standard number: +32 22999696,

— via the following form: european-union.europa.eu/contact-eu/write-us_en.

FINDING INFORMATION ABOUT THE EU

Online

Information about the European Union in all the official languages of the EU is available on the Europa website (european-union.europa.eu).

EU publications

You can view or order EU publications at op.europa.eu/en/publications. Multiple copies of free publications can be obtained by contacting Europe Direct or your local documentation centre (european-union.europa.eu/contact-eu/meet-us_en).

EU law and related documents

For access to legal information from the EU, including all EU law since 1951 in all the official language versions, go to EUR-Lex (eur-lex.europa.eu).

Open data from the EU

The portal data.europa.eu provides access to open datasets from the EU institutions, bodies and agencies. These can be downloaded and reused for free, for both commercial and non-commercial purposes. The portal also provides access to a wealth of datasets from European countries.

Science for policy

The Joint Research Centre (JRC) provides independent, evidence-based knowledge and science, supporting EU policies to positively impact society

EU Science Hub
joint-research-centre.ec.europa.eu

@EU_ScienceHub

EU Science Hub - Joint Research Centre

EU Science, Research and Innovation

EU Science Hub

@eu_science

Publications Office
of the European Union