

# A Methodology for the Evaluation of Innovation Incentive Programs on Firm Performance\*

**PROSPER**

Universidade Católica Portuguesa

November 30, 2022

## **Abstract**

This paper provides a comprehensive overview of empirical methods that can be used to evaluate the impact of innovation programs in Portugal. It also summarizes existing evidence on the outcomes of these policies for firms and reviews the information available in the main firm-level microdatasets. It provides concrete evidence on how to link and leverage these datasets for the evaluation of the effects of innovation policies on R&D and firm performance. The paper discusses the main advantages and drawbacks of the application of the different evaluation methods to the case of Portugal and how they could be used to improve program impact. Finally, it provides examples of evaluations done in other countries that can guide policy evaluation.

**Keywords:** VET, Education, Portugal

---

\*[joana.silva@ucp.pt](mailto:joana.silva@ucp.pt). We remain responsible for any errors. We thank POAT for financial support.

# Contents

<b>1</b>	<b>Introduction</b>	<b>4</b>
<b>2</b>	<b>Innovation Incentives: Policies and Outcomes</b>	<b>6</b>
2.1	Promoting Firm Investment . . . . .	6
2.2	Providing Incentives to Innovate . . . . .	7
2.3	Capturing Global Talent . . . . .	7
<b>3</b>	<b>Portuguese Data Setting</b>	<b>8</b>
3.1	Sistema de Contas Integradas da Empresa . . . . .	8
3.2	Comércio Internacional . . . . .	9
3.3	CIS . . . . .	10
3.4	IUTICE . . . . .	11
3.5	SIFIDE . . . . .	11
<b>4</b>	<b>Methods</b>	<b>13</b>
4.1	Key Concepts - Counterfactuals . . . . .	13
4.2	Randomized Controlled Trials . . . . .	15
4.3	Difference-in-differences . . . . .	26
4.4	Instrumental Variables . . . . .	33
4.5	Regression Discontinuity Design . . . . .	37
<b>5</b>	<b>Conclusion</b>	<b>43</b>

## List of Figures

1	Competition Stages and Treatment Assignment. Source: McKenzie (2017)	20
2	Baseline Characteristics and Balance of Experimental Sample. Source: McKenzie (2017) . . . . .	21
3	Impact on Employment, Firm Size and Innovation. Source: McKenzie (2017)	23
4	Impact on Business Sales and Profits. Source: McKenzie (2017) . . . . .	24
5	Impact on Capital. Source: McKenzie (2017) . . . . .	24
6	The Difference-in-differences method for causal inference . . . . .	27
7	Impact of the Intervention according to a Difference-in-differences approach	28
8	Baseline Results from diff-in-diff specification. Source: Guceri and Liu (2019) . . . . .	31
9	Average R&D Spending across Groups, Relative to 2007 R&D Spending. Source: Guceri and Liu (2019) . . . . .	32
10	Intuition behind the IV Approach . . . . .	33

11	User-cost elasticity of firm R&D intensity. Dependent variable: $\Delta$ (qualified R&D / sales). Source: Rao (2016) . . . . .	38
12	Pre-Assignment Mean-differences between Untreated and Treated Firms. Source: Bronzini and Iachini (2014) . . . . .	41
13	Baseline Results, Effect of the Program on Investment. Source: Bronzini and Iachini (2014) . . . . .	42

# 1 Introduction

Every year, governments around the world invest millions of euros in innovation incentive programs, with this type of policy gaining substantial prominence in recent years. The effects of investing in innovation have long been documented, with benefits ranging from firm competitiveness and productivity enhancement to economic growth at the aggregate level (Jones and Williams, 2000). However, there is a lot of uncertainty regarding the outcomes of these programs (Köhler et al., 2012; Mitchell et al., 2020). Even policies that were put in place in the past may have different effects than before if the economic context is different, or even when seemingly unrelated circumstances change in the meantime.

R&D is at the center of public debate in Portugal, where the government has recently announced the allocation of more than two billion euros to innovation in the context of the Recovery and Resilience Program. This calls for accurate and solid evaluation that can track processes and measure outcomes. What are the characteristics of the firms that benefit from this type of policy? And what are the results of these incentives in the composition of the workforce and growth prospects of each firm?

Despite the relevance of these policies in Portugal, there is no consistent methodology for evaluating public policies that focuses on real impact-based metrics, comprising both the short-run and the long-run. Indeed, the methods employed by public evaluators to measure results are extensively based on expenditure measurement and quantitative measures, contrasting vividly with the innovative tools and rigorous methods commonly used in the academic environment.

Naturally, the adoption of these methodologies largely depends on the existence of complete and quality information. In fact, the Portuguese information and statistics system is one of the best in the world, comprising relevant and extensive data that, when well processed and articulated, is capable of revolutionizing the way in which public policies are carried out and how results are monitored and measured. Overlooking this information implies missing the opportunity to fully understand the impact of each project and the mechanisms that explain its success or failure, ultimately leading to lower efficiency in the allocation of public funds.

This project contributes to the discussion of accurate and relevant policy evaluation by providing an overview of cutting-edge empirical methods relevant to policy evaluation and reviewing the main relevant Portuguese datasets. We provide detailed information on the actual implementation of each method and discuss real examples of applications to innovation policies in other countries. Furthermore, we include an extensive overview of the existent data for Portugal, discussing how to leverage it and how to take advantage of the interconnectability of the data.

This document is organized as follows. Section 2 provides a guide into innovation

policy, i.e. its rationale, goals, and (documented) effects, particularly through the use of examples from incentive programs implemented in other countries. Section 3 describes the existing databases in Portugal with relevant information for policy evaluation in the field of innovation. Section 4 presents the counterfactual evaluation methods, with a rigorous base in econometrics and intuitive interpretation, as well as documented shortcomings. To motivate each method, we present a paper employing the design to the evaluation of a policy in R&D and innovation. Finally, section 5 concludes.

## 2 Innovation Incentives: Policies and Outcomes

In today's globalized world, innovation plays a major force behind competitiveness. It is the foundation for long-term prosperity and economic success. The performance of economies is becoming more dependent on information and services, where investment in intangible assets is crucial and where governments look to innovation-induced development as a spark to re-ignite growth. To foster a supportive policy climate, governments assist the development and adoption of innovation through a variety of avenues (regulatory policy, sound tax and financial system). A cogent and well-planned innovation strategy must include tax policies to promote overall long-term and sustainable growth. Nevertheless, tax credit incentives must be used with caution since they increase the complexity of the system, compromising transparency.

As far as Portugal is concerned, the country has more than 500 tax expenditure regimes aimed at firms and individuals. On the firm side, there are four key tax benefits targeted at fostering investment and innovation. On the individual side, the most well-known type of tax benefits are those on IRS. Yet, other than IRS, there are individual tax benefits with effects on firms through labor composition, namely the non-habitual resident tax regime which aims to attract foreign talent.

### 2.1 Promoting Firm Investment

In Portugal, there are two tax incentives to promote investment in the country: Regime Fiscal de Apoio ao Investimento (RFAI) and Dedução por Lucros Retidos e Reinvestidos (DLRR).

The RFAI provides tax deductions for investments done in certain regions and in certain sectors. The goal is to boost investment and provide incentives for firms to locate in certain regions and sectors. However, international evidence shows that tax deductions are generally not sufficient to attract major flows of investment. Investors often emphasize the relative nonentity of the tax system in investment decisions compared with other considerations (tax base, tax rates, country's economic situation, stability).

The DDLR allows for the deduction of the retained earnings that are reinvested, in relevant investments, from the collection of the IRC. The objective is to provide incentives for growth and capitalization at the left tail of the corporate tissue. Evidence highlights that tax incentives and deductions for SMEs typically produce a mixed bag of outcomes, according to data on tax relief. In fact, the vast majority of EU regulations don't offer SMEs much relief. Tax incentives in Portugal significantly help medium-sized firms, while they scarcely help small firms. The issues in the SME sector should be addressed directly rather than using tax incentives, which are an ineffective method to do so.

## 2.2 Providing Incentives to Innovate

In Portugal, the Sistema de Incentivos Fiscais à I&D Empresarial (SIFIDE II) and the Patent Box are used to provide incentives to innovate.

SIFIDE II, which is one of the most generous tax credit systems among OECD countries, is a hybrid tax credit system on R&D expenses. Since 2011, a significant number of SMEs receive SIFIDE II. Portuguese evidence on SIFIDE II points out a positive impact of the program on R&D investments and confirms input additionality. However, international evidence remains mixed on this subject: some studies confirm additionality, whereas others report crowding-out effect. Results are heterogeneous and may depend on features of tax credits and also on firm-specific characteristics.

On the other hand, the patent box regime provides a 50% tax exemption on income derived from intellectual property, including income generated from patents, designs and models that are the result of internal R&D. In 2019, 703 patent applications were made in Portugal, a significant increase from the 81 applications in 2000. The goal is to promote R&D, firm innovation and intellectual capital formation. However, international evidence provides mixed results on Patent Boxes: in UK, patent boxes increase investment by 10% in firms who benefited from the program; on the other hand, studies using data from the Netherlands show that there is indeed a positive effect on R&D investment, even though such policy has been argued to lead to tax shifting. There is also evidence that moving their patenting location won't have a meaningful impact on R&D spending if they choose the most forgiving patent box system.

## 2.3 Capturing Global Talent

NHR tax regime is a special tax system applied to those who become tax residents in Portugal after being non-tax residents for the preceding five years. It provides a flat rate of 20% to individuals who perform "high-value" activities (such as doctors, engineers, managers, etc.). The aim is to attract international talent and to promote immigration in Portugal, in order to rise consumption and investment.

According to recent data, the number of program participants has significantly increased, particularly after 2014, with the primary beneficiaries being from France, the United Kingdom, and Italy.

The number of recipients more than tripled between 2014 and 2018.

Beneficiaries work in high-skilled occupations in substantial numbers, the bulk of whom are managers and directors.

## 3 Portuguese Data Setting

This section covers how to leverage Portuguese microdata to evaluate firms' incentives to innovate.

### 3.1 Sistema de Contas Integradas da Empresa

Sistema de Contas Integradas das Empresas (SCIE) is a dataset compiled by the Instituto Nacional de Estatística (INE) providing firms' accounting data. It is produced on an annual basis and covers almost all Portuguese firms. SCIE's power stems from the fact that it aggregates several microdata from various sources in order to provide the most accurate characterization of the firm's financial behavior. This makes SCIE the primary source for firm-level financial information in Portugal.

SCIE includes firms that are societies and sole proprietorships, as well as independent workers who participated in any activity of production of goods and services while living in Portugal during the reference year. Firms must also be "economically active" to be included, meaning that during the reference period firms must have had some volume of sales or expenditures. Nonetheless, and considering the aforementioned definition, some entities are excluded from SCIE:

- Branches of the public administration, either central or local;
- Non-profit associations;
- Firms registered in Zona Franca da Madeira;
- Financial firms, which include banks and insurance firms.

SCIE provides a detailed income statement for each firm, which is identified by a fictitious ID, i.e., an anonymized rearranged ID constructed from the true fiscal number, in a panel form. This ID not only allows SCIE to be merged with all of the other datasets mentioned in this chapter (as well as with Quadros de Pessoal (QP), where each of these firms' payroll can be checked), but it also allows the firm to be tracked over time while knowing its sector, geographical location, founding year, and financial indicators. Key indicators reported in SCIE include the value of production, profits, labour costs, expenses with materials and external services, taxes, several types of investment and disinvestment. SCIE also provides simplified accounting information for some firms at the establishment level since 2008.

As mentioned before, SCIE aggregates information from several administrative sources, meaning that the variables mentioned are not available at all times and are collected by different institutions:



- Ficheiro de Unidades Estatísticas (FUE), which is organized by INE, feeds SCIE with information about existing firms in Portugal (both societies and sole proprietorships);
- A protocol between INE and Autoridade Tributária (AT), organized by the Ministry of Finance, where SCIE gets the accounting information about sole proprietorships and independent workers;
- Informação Empresarial Simplificada (IES), under the responsibility of the Ministry of Finance, the Ministry of Justice, the Bank of Portugal and INE, feeds SCIE with financial information about societies, making it its main source;
- Quadros de Pessoal (QP), collected by the Ministry of Solidarity, Employment and Social Security, provides SCIE with complementary information for establishment-level data.

This type of accounting microdata with basic firm-level financial information has undergone some transformations over time. Similar data has been collected in Portugal since 1994 through a dataset known as Inquérito Anual às Empresas (IEH)<sup>1</sup>. This dataset was available until 2003. Since 1994, various measures have progressively deepened the available data. In 2007, IEH was substituted by IES<sup>2</sup>. In 2010, Sistema Normalização Contabilística (SNC) launched the SCIE series. The changes mandated by SNC were applied retroactively until 2004, which is why SCIE is available since then (although the decision is from 2010).

SCIE can be used to see the financial effects of innovation and innovation policies on firms' performance through various lenses, like production, sales, profits, or any other accounting indicator.

## 3.2 Comércio Internacional

Comércio Internacional (CI) is a dataset produced monthly by INE containing firms that exported or imported products. CI provides an ID for the firm that allows for matching with other microdata and track firms through time. This dataset also provides detailed information about the country of origin or destination of each transaction, what good was traded, in what quantity, at what price, from or through what port or airport, whether Portugal was just an intermediary country, and what country originally produced that good. Because of the level of detail of this information at the firm level, there are nearly 10 million observations per year, despite the fact that Portugal has approximately 300,000 firms.

---

<sup>1</sup>Regulamento (CE, Euratom) N<sup>o</sup> 58/97 do Conselho, de 20 de dezembro de 1996.

<sup>2</sup>Decreto-Lei n<sup>o</sup> 8/2007.

However, CI has a level of complexity that researchers need to be aware of. CI actually joins two datasets for which INE is responsible: Extrastat and Intrastat.

Intrastat, compiled since 1993, when the Schengen Agreement was signed, concerns trade between Portugal and countries in the Schengen Area. The data is collected through a mandatory monthly survey for firms above two thresholds: one of assimilation and another of simplicity. The threshold of assimilation comprises the boundaries of monthly intra-EU imports and exports after which firms are legally mandated to respond. The threshold changes every year. The maximum value for the threshold was verified in 2009, when only firms with either more than 400 thousand euros in intra-imports or 550 thousand euros in intra-exports were mandated to answer. The minimum was 39 904 for imports, in 1993, and 80 thousand for exports, in 2006.

On the other hand, the threshold of simplification allows firms below it to respond in a less detailed manner to the survey. Below this threshold, firms can only give detailed product information for their top 10 traded goods and are exempt from disclosing both quantities and the nature of the transaction. The threshold also varies on a yearly basis. The maximum values were 550 thousand for intra-imports and 750 thousand for intra-exports, both in 2009. The minimum values were 59 856 in 1993 for both.

Extrastat, compiled since 1988, concerns, since 1993, trade between Portugal and countries outside the Schengen Area. It is built with "Documentos Únicos" (DU), which are forms that firms must fill out and deliver to customs. Here, any firm that does any export or import to countries outside of Schengen Area is reported, along with the good, price, quantity, and respective countries. Thus, there are no thresholds for a firm to be reported.

CI must be used with caution, given that there may be some sample selection issues, depending on the economic problem. Small firms are expected to be excluded by the threshold of assimilation (or not respond in a detailed manner due to the threshold of simplicity). Even worse, the thresholds change throughout the years, making the sample selection mechanism time-inconsistent. Notwithstanding, analyzing trade and innovation is still possible with the right setting. This is because INE ensures that the thresholds are set in such a way that at least 97% of all the trade between EU member states is in CI, as well as 6% of total commercial trade (also considering trades within Portugal).

### 3.3 CIS

Inquérito Comunitário à Inovação (CIS) is the main dataset INE compiles focusing on innovation and R&D. It has a firm ID that allows it to be connected to the other data in this chapter, and it is also microdata. However, it is not administrative data, but a survey. The survey draws observations from the Portuguese population of active firms with more than 10 employees.

The CIS survey is biennial (conducted every two years). The units responsible for this data are INE, Direção-Geral de Estatísticas da Educação e Ciência, and Direção de Serviços de Estatística da Ciência e Tecnologia e da Sociedade da Informação. Because it follows methodological recommendations from Eurostat, it is fully compatible with its sibling studies in European countries, making this dataset very attractive for researchers wanting to compare Portugal to other nations.

CIS covers extremely detailed variables about both the investment in innovation and the practical use of those innovations, as well as their efficacy. This includes: sales or spending in intellectual property, consulting firms, trademarks, patents, and copyright; what were the reasons for the spending in R&D and reasons for not spending in R&D; the efficacy of previous spending in R&D; expectations of future investment in R&D; and many more.

### **3.4 IUTICE**

While CIS surveys R&D and innovation, the Inquérito à Utilização de Tecnologias da Informação e da Comunicação (IUTICE) dataset does a similar task with information and communications technology (ICT). This dataset is produced annually by INE and provides comprehensive information on the adoption of technologies by Portuguese firms. It has been available since 2004.

The survey of IUTICE is done in a stratified manner. IUTICE contains a yearly census of all Portuguese firms with more than 250 employees and probability-weighted samples of other smaller firms. This makes the sample smaller than 6 thousand firms per year. Financial institutions have been excluded from this survey since 2014.

Variables reported include: sales from online channels, the difficulty of hiring ICT staff, number of ICT staff that the firm has, usage of robots (and for what end), security problems related to ICTs, internet's maximum speed and tasks that can be done on the firm's internal server (if it exists). Although IUTICE does not report on variables related to a firm's investment in innovation, it's extremely useful to study the adherence to already-existing technologies and the respective firm's outcomes from it.

### **3.5 SIFIDE**

Finally, Sistema de Incentivos Fiscais à I&D Empresarial (SIFIDE). SIFIDE is an administrative microdata related to the tax credit system mentioned in section 2.1. The data contains information about firms that applied for SIFIDE support, from 2006 until today, whether it was granted or not. Because the SIFIDE program was started in 2006, the dataset covers the entire lifespan of the program.

SIFIDE has, for those firms, a firm ID crossable with every other dataset here in this

section, their spending on R&D, tax credits requested and tax credits provided.

This dataset can be crucial to studying the effect of innovation since it indicates which firms got financial support to innovate and which didn't; aside from being obviously useful to conduct a policy analysis to evaluate SIFIDE's capacity to incentivize innovation.

## 4 Methods

In this chapter, we will present five methods for policy impact evaluation, based on modern econometric techniques and which have been widely employed in the assessment of the effects of policies all over the world, by governments, other public entities, world-renowned institutions and private firms. We will begin the discussion with a section on the importance of counterfactual impact evaluation, stating the key concepts we will employ and mention throughout the chapter.

It is important to bear in mind that the choice of an impact evaluation method depends to a great extent on aspects such as the operational characteristics of the program being evaluated, available resources, eligibility criteria for selecting beneficiaries, and timing for program implementation.

A methodology for policy impact evaluation must give answer to the specific cause-and-effect question: *What is the impact of a program on an outcome of interest?*. This basic question incorporates an important causal dimension. The focus is solely on impact, i.e. the changes directly attributable to a program.

The first idea that could come to mind is to use what is called an *event study*: Comparing an outcome for a group of units (individuals, firms, countries) before and after an event we believe might affect the outcome. In the context of policy evaluation, this means comparing the outcome of interest for the firms that are part of a program after they enter the program with the outcome before, possibly accounting for a time trend.

The main drawback of this approach is, however, the fact that the difference in means pre-intervention and post-intervention cannot be interpreted as a causally valid estimate of the effect of the policy, since other factors influencing the outcome may serve as *confounding factors*.

Therefore, although cause-and-effect questions are common, answering them accurately is quite challenging. To establish causality between a program and an outcome, we employ impact evaluation methods to rule out the possibility that any factors other than the program of interest explain the observed impact.

### 4.1 Key Concepts - Counterfactuals

We can think of the impact of a program as the difference in outcomes for the same unit (person, household, community, firm, and so on) with and without participation in a program. Yet, it is impossible to measure the same unit in two different states at the same time - at any given moment in time, a unit either participated in the program or did not participate. Therefore, we need to rely on a comparison between two units that we believe would have behaved in a similar way in the absence of treatment and compare

the outcomes of interest.

Since no perfect clone exists for a single unit, the key to estimating the counterfactual for program participants is to move from the individual or unit level to the group level. We can rely on statistical properties and methods to generate two groups of units that, if their numbers are large enough, are statistically indistinguishable from each other at the group level. The group that participates in the program is known as the **treatment group**. The statistically identical group is known as **comparison** or **control group**. This group remains unaffected by the program, and allows evaluators to estimate the counterfactual outcome, that is, the outcome that would have prevailed for the treatment group had it not received the program. If the two groups are identical, with the sole exception that one group participates in the program and the other does not, then we can be sure that any difference in outcomes must be due to the program.

A valid comparison group must:

- have the same characteristics, on average, as the treatment group in the absence of the program;
- remain unaffected by the program;
- react to the program in the same way as the treatment group, if given the program.

Finding such comparison groups is the core of any impact evaluation, regardless of what type of program is being evaluated. Indeed, without a comparison group that yields an accurate estimate of the counterfactual, the true impact of a program cannot be established. In addition to this, the degree of comparability between treatment and comparison groups is central to the and is therefore fundamental to assessing a program's causal impact.

Failing to estimate the correct control group threatens the evaluation's *internal validity*. Two commonly used methods to estimate the counterfactual, the *before-and-after* and *enrolled-and-non-enrolled* comparisons, fail to comply with the aforementioned desirable characteristics for a good counterfactual:

- Before-and-after comparisons: They compare the outcomes of the same group before and after participating in a program. This is not a valid comparison because it fails to account for other time-varying factors other than treatment that may influence the outcomes.
- Enrolled-and-non-enrolled (or self-selected) comparisons: They compare the outcomes of a group that chooses to participate in a program with those of a group that chooses not to participate. Due to selection bias and differences in unobservable characteristics between groups we cannot ensure that the two groups are similar and that they would react to the program in the same way.

In the next sections, we summarize some of the methods which can be employed in a policy evaluation context, referring to their econometric estimation, intuition and illustrative real-life applications in an R&D policy setting.

## 4.2 Randomized Controlled Trials

A randomized controlled trial (RCT) is an experimental method of impact evaluation in which all eligible units in a sample (for example, an individual, household, business, school, hospital, community, etc.) are randomly assigned to treatment and control groups - the treatment group receives or participates in the program being tested, while the control group does not. Given a sufficiently large number of units, an RCT ensures that the control and treatment groups are equal in both observed and unobserved characteristics, thus ruling out selection bias. Therefore, the only difference between the treatment and control groups is their participation in the intervention itself, and the difference in their outcomes represents the impact of the intervention or program ([World Bank \(n.d.\)](#)).

RCTs are considered the gold standard of impact evaluation. This method employs a random process to decide who is granted access to the program and who is not. Under randomized assignment, every eligible unit has the same probability of being selected for treatment by a program. Therefore, the randomized assignment process in itself will produce two groups that have a high probability of being statistically identical, provided that the number of potential units to which the treatment is applied is sufficiently large - hence why this method produces an excellent estimate of the counterfactual.

With the baseline data from the evaluation sample, it is possible to test this assumption empirically and verify that in fact there are no systematic differences in observed characteristics between the treatment and control groups before the intervention. Then, after the intervention, if there are observable differences in outcomes between the treatment and control groups, we will know that those differences can be explained only by the introduction of the program, since by construction the two groups were identical at the baseline and are exposed to the same external environmental factors over time.

### Implementation

The implementation of this method can be described in the following steps ([White et al. \(2014\)](#)).

Firstly, one should define the units that are eligible for the program, while verifying that the eligible population is greater than the number of program slots available. Once this is done, it will be necessary to compare the size of the group with the number of observations required for the evaluation. The size of the evaluation sample is determined through power calculations and is based on the types of questions the evaluation aims at

answering. To conduct power calculations and estimate the required sample size for an evaluation, evaluators usually use assumptions regarding the expected effect size, the statistical significance level and the intracluster correlation (for cluster RCTs). The intracluster correlation is a descriptive statistic between 0 and 1 that indicates how strongly the groups (e.g., households) or the individuals in the cluster resemble each other. The higher the intracluster correlation, the higher the required sample size. In cluster RCTs, there is usually a greater increase in statistical power when the number of clusters is increased than when the number of individuals or groups within a cluster is increased.

The second step is to form the treatment and control groups from the units in the evaluation sample through randomized assignment. Random assignment should not be confused with random sampling. Random sampling refers to how a sample is drawn from one or more populations. Random assignment refers to how individuals or groups are assigned to either a treatment group or a control group. RCTs typically use both random sampling (since they are usually aiming to make inferences about a larger population) and random assignment (an essential characteristic of an RCT). Randomization is typically done at the level at which the program is implemented. As the level of randomized assignment gets lower, the chances increase that the control group will be inadvertently affected by the program. On the other hand, when the level of the randomized assignment is higher or more aggregate, it becomes increasingly difficult to perform an impact evaluation because the number units can be insufficiently large to yield balanced treatment and comparison groups.

Randomization can be done at the **individual level**, where individual units are assigned to a treatment or control group ([World Bank, n.d.](#)). Alternatively, in **cluster randomization**, clusters of units rather than the units themselves are randomly assigned to treatment and control groups (i.e. cohort, village). Clustered RCTs are the preferred type of RCT when the intervention is by definition applied at the cluster rather than the individual level (i.e. an intervention targeted towards schools or health facilities in a given setting). The statistical power in cluster RCTs is typically lower than that for individually randomized trials, since outcomes within clusters are typically somewhat similar to each other. This means that the number of clusters in a cluster RCT, rather than the number of individuals who participate, is most relevant to the statistical power of the study. Cluster RCTs are often more expensive than individually-randomized RCTs. However, cluster RCTs provide administrative convenience, reduce ethical concerns, and avoid treatment group contamination. Additionally, we can consider **phase-in randomization**, where the roll-out of the intervention is randomized and every unit or cluster in the population of interest will get the program eventually. Phase-in designs are usually used at the cluster-level but may also be applied at the individual-level. Randomized phase-ins are easily



applied to project implementation schedules, as roll-outs typically happen over multiple years. These also reduce concerns of inequity and provide incentives to maintain contact. However, for control participants, phase-in designs could change present actions through setting expectations of future change. Moreover, phase-in designs complicate estimating long-run effects since once the intervention is fully rolled out, no control group remains. Long-run analyses can still examine differences between groups with degrees of exposure.<sup>9</sup>

It is possible to randomly assign population groups to the treatment and control groups in several ways, including:

- Simple randomization – Individuals or sites are listed and then assigned to the treatment and control groups using random numbers, for example, issued by a random number generator.
- Matched pair randomization – Individuals or clusters are grouped into pairs based on having similar observable characteristics. One unit in each pair is randomly assigned to the treatment group and the other to the control group. This initial matching helps to ensure balance and reduces the required sample size.
- Stratified random assignment – For key variables likely to influence results, participants are divided into groups (strata) and then randomization is conducted for each group. This ensures an equivalent distribution of key variables across the treatment and control groups.

There are two particular types of risks to consider when choosing the level of assignment: spillovers and imperfect compliance/crossovers. Spillovers occur when the treatment group directly or indirectly affects outcomes in the comparison group (or vice versa). Spillovers may be physical, behavioral, informational, market or general equilibrium. Imperfect compliance occurs when some members of the comparison group participate in the program, or some members of the treatment group do not. In either of these situations, the validity of the control group is compromised because some control units receive treatment. In both cases, the comparison group no longer serves as a counterfactual. Nevertheless, carefully considering the level of randomized assignment, can minimize the risk of spillovers and imperfect compliance.

After the randomized assignment of groups, and before the program begins, baseline data on the population of interest should be employed to verify that there are no systematic differences in observed characteristics between the treatment and control units, that is, verifying the balance between both groups.

Lastly, once a random evaluation sample has been selected and treatment has been assigned in a randomized fashion, it is quite straightforward to estimate the impact of the

program. After the program has run for some time and at the end of its implementation, outcomes for both the treatment and comparison units will need to be measured. The impact of the program is simply the difference between the average outcome ( $Y$ ) for the treatment group and the average outcome ( $Y$ ) for the comparison group.

## Validity

A crucial discussion to consider when applying RCTs is the question of Validity. *Internal validity* means that the estimated impact of the program is net of all other potential confounding factors, that is, the comparison group provides an accurate estimate of the counterfactual, so that we are estimating the true impact of the program. On the other hand, an RCT must also establish *External validity*, which allows to conclude that the evaluation sample accurately represents the population of eligible units. Hence, the results of the evaluation can then be generalized and extrapolated to the population of eligible units.

An impact evaluation can produce internally valid estimates of impact through randomized assignment of treatment. However, if the evaluation is performed on a nonrandom sample of the population, the estimated impacts may not be extrapolated to the population of eligible units, jeopardizing external validity. Alternatively, if the evaluation uses a random sample of the population of eligible units, but treatment is not assigned in a randomized way, then the sample would be representative, but the comparison group may not be valid, thus jeopardizing internal validity.

Thus, it is essential to consider the steps outlined above for randomized assignment of treatment, in order to ensure both the internal and the external validity of the impact estimates - the random selection of a sample establishes external validity and the randomized assignment of treatment as an impact evaluation method establishes internal validity.

## Shortcomings

RCTs face a range of ethical and practical concerns.

On the ethical side, there are particular ethical concerns around RCTs that relate to their experimental nature and which make it important for participants in the trial to be consulted and their wishes identified and addressed, and for the associated risks and benefits to be balanced.

The ethical concerns around experimentation become even more striking in the case of RCTs that involve a control group that does not receive any intervention. The potential for disadvantage to consequently occur makes it very important that randomization is

a transparent process, especially when randomizing at the individual level. It is the evaluator's responsibility to ensure that no tensions exist between the treatment and control groups. One of the ways of mitigating this possibility is by clearly explaining the purpose of randomization.

Another ethical concern surrounds the need for an RCT in the first place. When there is no reasonable doubt about the benefits and cost-effectiveness of a programme, then there is no need for an in-depth evaluation (of any kind) and impact monitoring may be more appropriate to assess whether the programme continues to have the intended results over time. If there are questions about a programme's effectiveness, and only limited resources available for its implementation, however, it may be considered most ethical to randomly assign the participants to the programme – with the intention of rolling out the programme to the whole population if it is found to be effective.

It is also crucial to be sensitive about the data collection involving the control group. Evaluators need to take into consideration that they are using the time of non-recipients appropriately. It is sensible to compensate the respondents in a survey for their time, although it should be done in such a way that it does not affect the results.

On the other hand, there are also practical considerations surrounding the application of an RCT design. Regarding the logistics, power calculation might demand vast sample sizes, which require increasingly more resources from the investigators. Additionally, validity requires multiple sites, which can be difficult to manage. Long trial run time may also result in the loss of relevance as practice may have moved on by the time the trial is published.

***Identifying and Spurring High-Growth Entrepreneurship: Experimental Evidence from a Business Plan Competition* by [McKenzie \(2017\)](#)**

In this 2017 paper published in the *American Economic Review*, David McKenzie utilizes a large-scale national business plan competition in Nigeria in an RCT setting, in order to assess if there are potential high-growth entrepreneurs, and the role public policy may play in helping identify them and facilitate their growth.

The author starts by identifying that in Nigeria 99.6 percent of firms have fewer than ten workers. As highlighted in the literature, the move away from self-employment toward wage employment in firms of larger sizes is a key aspect of the development process, which raises the key policy questions of whether there are constrained entrepreneurs in developing countries with the ability to grow a firm beyond this 10-worker threshold.

To answer this question he exploits the YouWiN! competition, launched in late 2011, an initiative which attracted in its first year of implementation almost 24,000 applications aiming to start a new business or expand an existing one.

## Implementation and Randomized Assignment

The competition setting is as follows:

- The top 6,000 applications were selected to receive an intensive training course. Then, from these applications, winners were selected to win around US\$50,000 cash prize, paid out in tranche payments conditional on achieving basic milestones.
- The top-scoring plans overall and within region were chosen as winners automatically.
- 729 additional winners were randomly selected from a group of 1,841 semifinalists, providing experimental variation from US\$34 million in grants that enables causal estimation of the program's impact.

Figure 1 below provides a detailed schematic representation of the competition stages and treatment assignment.

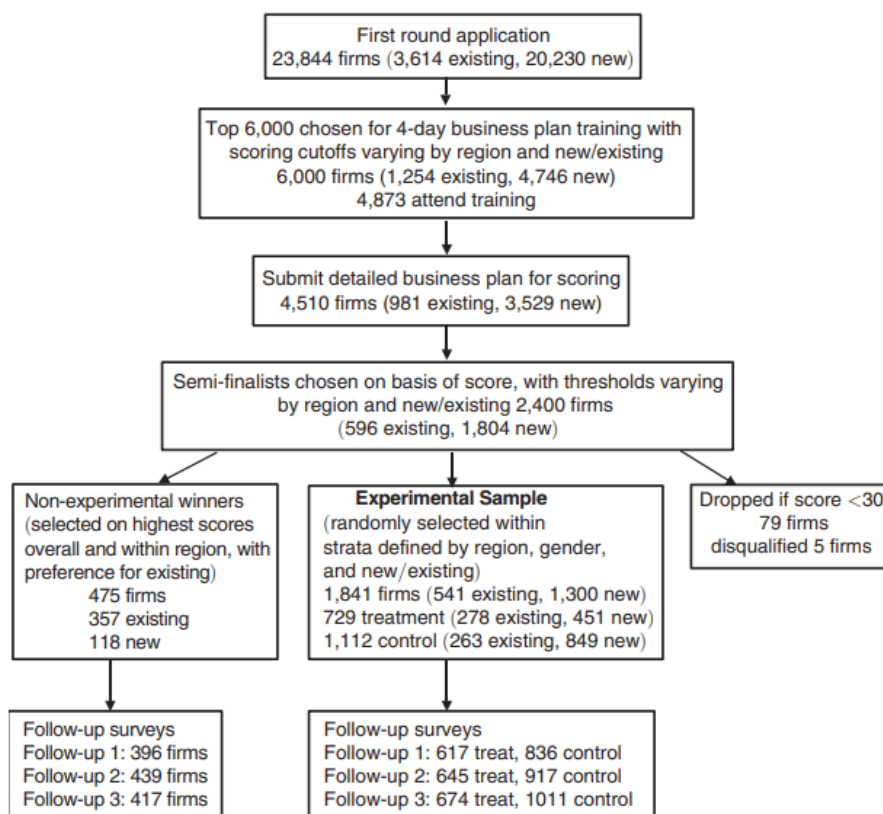


Figure 1: Competition Stages and Treatment Assignment. Source: McKenzie (2017)

As far as the randomization algorithm is concerned, random selection of the ordinary winners was done considering these main reasons: firstly, from an operational point of view, given the large-scale of a competition of this nature; secondly, to address concerns

that programs get captured by individuals with certain political or ethnic ties; and lastly, random assignment enabled rigorous measurement of the program's impacts. This constitutes an ideal setting to the application of an RCT research design for evaluation.

Finally, after quality checking efforts, this allocation yields 1,841 firms in the ordinary winner pool, such that the **treatment group** is consisted of 729 firms randomly assigned to treatment. The **control group** consists of the 1,112 firms randomly assigned to the control group (after the random replacement). In terms of the impact evaluation, this will be handled through assignment to treatment analysis.

In line with the steps outlined in the sections above, to establish the validity of the experiment the author performs balance checks between the firms attributed to treatment vs. control, to verify there are no significant differences in characteristics across both groups. The results of this analysis can be seen in Figure 2

	Existing firms			New firms		
	Non-experimental winners	Treatment group	Control group	Non-experimental winners	Treatment group	Control group
<i>Applicant characteristics</i>						
Female	0.17	0.18	0.17	0.19	0.17	0.18
Age	32.5	32.0	31.8	30.1	29.3	29.6
Married	0.60	0.50	0.56	0.42	0.34	0.36
High school or lower	0.10	0.13	0.12	0.06	0.11	0.10
University education	0.71	0.63	0.67	0.79	0.69	0.71
Postgraduate education	0.12	0.08	0.12	0.13	0.05	0.06
Lived abroad	0.14	0.10	0.11	0.18	0.06	0.09
Choose risky option	0.59	0.56	0.52	0.63	0.57	0.55
Have internet access at home	0.68	0.57	0.61	0.60	0.47	0.48
Own a computer	0.94	0.87	0.88	0.92	0.84	0.86
Satellite dish at home	0.74	0.67	0.71	0.64	0.68	0.64
Freezer at home	0.64	0.57	0.61	0.63	0.51	0.55
<i>Business characteristics</i>						
Crop and animal sector	0.14	0.16	0.16	0.22	0.22	0.22
Manufacturing sector	0.28	0.28	0.26	0.23	0.28	0.24
Trade sector	0.05	0.06	0.05	0.06	0.04	0.04
IT sector	0.14	0.15	0.14	0.04	0.07	0.06
First round application score	59.0	57.2	56.6	59.9	59.9	59.9
Business plan score	61.7	45.8	45.4	74.4	53.7	55.5
Number of workers	9.11	7.35	7.73			
Ever had formal loan	0.11	0.06	0.09			
<i>Sample size</i>						
	357	278	263	118	451	849
Joint orthogonality test: treatment versus control		0.920			0.884	
Joint orthogonality test: non-experimental versus treatment	0.000			0.000		
Joint orthogonality test: non-experimental versus treatment (no score)	0.012			0.000		

Figure 2: Baseline Characteristics and Balance of Experimental Sample. Source: McKenzie (2017)

This establishes no significant differences between groups, ensuring the randomized sample is balanced. This exercise is also useful for showing some basic characteristics of the experimental sample. For instance, the average existing firm owner is male, aged 32, with 4 years of business experience and running a business with a median of 5 workers. On the other hand, new applicants are slightly younger, with an average age of 29, 70

percent have university education, and they come from relatively well-off households, with 85 percent having a computer at home and two-thirds having a satellite dish.

Regarding the data collection process, 3 follow-up annual surveys were employed to track the evolution of the winners, followed by a fourth, longer-term follow-up. The first follow-up survey took place approximately one year after individuals had first applied to the program, the second survey took place approximately two years after application and just as firms had received their last tranche payments, and the third follow-up survey took place three years after application, and between 12 and 18 months after firms had received their last tranche payment from the program.

## Results

As far as the empirical estimation of the program impacts are concerned, the main approach used for evaluating the impact of the program is to use the randomized controlled trial (RCT) based on the random selection of ordinary winners from among the semifinalists. This is done separately for the new and existing business applicants, and involves regressions of the following form:

$$Outcome_i = \beta_0 + \beta_1 AssignTreat_i + \beta_2 Region * Gender_i + \epsilon_i$$

where  $AssignTreat_i$  denotes whether or not applicant  $i$  was randomly chosen as an ordinary winner from among the semifinalist experimental pool, and  $Region * Gender_i$  controls for the randomization strata.

The results for the outcomes employment (one of the main goals of the program), firm size and innovation can be seen in Figure 3 below. The results for firm performance outcomes such as profits and sales can be seen in Figure 4. Finally, the authors also perform an analysis of the impacts on firm capital are specified in Figure 5

The first column shows a positive impact on the employment status of the owner. Column 2 then considers total employment in the firm, the average control group firm among new applicants has 3.7 workers by the time of the third survey, with the treatment effect of 5.2 workers more than doubling this average. Impacts are larger in the second and third years once all the grants had been received than in the first year. Column 3 examines the extent to which winning the competition has enabled firms to surpass the ten worker threshold. Amongst new firm applicants, we see that only 11 percent of the control group had reached this size three years after applying, with treatment increasing this by 22.9 percentage points. Among existing firms, 17 percent of the control group were at this size after three years, with the treatment taking a further 20.6 percent to this level. Few firms have grown to the size of having 25 workers, but column 4 shows that by the third round the treatment has had a statistically significant 2.5–2.7 percentage point

	Own employment	Total employment	Firm of 10+ workers	Firm of 25+ workers	Innovation index
<i>Panel A. New firms</i>					
First follow-up	0.074 (0.025)	1.426 (0.732)	0.024 (0.020)	0.007 (0.008)	0.099 (0.019)
Second follow-up	0.128 (0.017)	6.012 (4.412)	0.288 (0.026)	0.022 (0.009)	0.270 (0.018)
Third follow-up	0.119 (0.018)	5.227 (4.469)	0.229 (0.028)	0.025 (0.011)	0.219 (0.019)
Control mean: first follow-up	0.787	3.618	0.083	0.010	0.225
Control mean: second follow-up	0.841	3.305	0.088	0.009	0.214
Control mean: third follow-up	0.831	3.773	0.114	0.014	0.181
Sample size: first follow-up	1,021	987	987	987	995
Sample size: second follow-up	1,181	1,159	1,159	1,159	1,071
Sample size: third follow-up	1,085	1,044	1,044	1,044	927
<i>Panel B. Existing firms</i>					
First follow-up	0.047 (0.019)	1.512 (0.795)	0.057 (0.041)	0.007 (0.019)	0.105 (0.029)
Second follow-up	0.066 (0.018)	2.556 (1.388)	0.215 (0.041)	0.009 (0.018)	0.126 (0.028)
Third follow-up	0.070 (0.022)	4.425 (0.673)	0.208 (0.040)	0.028 (0.015)	0.141 (0.029)
Control mean: first follow-up	0.938	6.852	0.212	0.032	0.390
Control mean: second follow-up	0.922	8.134	0.231	0.038	0.407
Control mean: third follow-up	0.906	5.571	0.170	0.014	0.341
Sample size: first follow-up	432	422	422	422	423
Sample size: second follow-up	505	500	500	500	458
Sample size: third follow-up	477	461	461	461	409

Figure 3: Impact on Employment, Firm Size and Innovation. Source: McKenzie (2017)

increase in this likelihood, relative to a control mean of only 1.4 percent.

The other stated objective of the program was to encourage innovation. This is measured as 12 different types of innovative activities aggregated into an index. The last column shows that winners are also innovating more. By the final survey round there is a 22 percentage point increase in innovative activities for experimental winners among new firms, and 14 percentage point increase for existing firms.

As we can see, the impacts are stronger in years 2 and 3 than in year 1, for both existing and new firms. For existing firms the impact on the level of sales is significant in years 2 and 3, while the impact on profits is only significant in the second year. For new firms, the impact on profits and sales is only statistically significant in the second year.

Finally, the author examines how winning affected the firm's use of capital. The use of both forms of finance (loans and equity investment) is very low, with fewer than 3 percent of the new firm applicants in the control group receiving either form of financing in a given year, and only 6 percent of existing firm applicants having received a formal loan in any given year, and fewer than 5 percent receiving equity investments.

The remaining columns, in contrast, show that the grants greatly increased the amount of capital in the winning firms. Treated firms have higher inventory levels, are more likely to have purchased business equipment, land, or buildings and have spent more on such purchases. The result is that the total capital stock of the treated firms is 3.5 million naira

	New firms				Existing firms			
	Truncated sales	Truncated profits	Inverse hyperbolic sine profits	Aggregate index of sales and profits	Truncated sales	Truncated profits	Inverse hyperbolic sine profits	Aggregate index of sales and profits
<i>Experimental impacts</i>								
First follow-up	36.160 (49.884)	-24.512 (26.330)	2.156 (0.369)	0.016 (0.047)	50.805 (85.662)	0.074 (49.416)	0.972 (0.373)	0.080 (0.070)
Second follow-up	297.783 (56.494)	69.061 (15.150)	4.154 (0.326)	0.298 (0.036)	346.304 (134.728)	69.234 (35.420)	2.183 (0.401)	0.237 (0.060)
Third follow-up	64.541 (92.338)	20.137 (21.635)	3.962 (0.346)	0.167 (0.042)	349.228 (143.729)	32.035 (40.956)	2.580 (0.464)	0.211 (0.070)
Control mean: first follow-up	271.467	167.705	6.583	-0.005	509.699	257.025	10.772	-0.045
Control mean: second follow-up	278.177	91.061	6.161	-0.096	660.535	206.305	9.646	-0.117
Control mean: third follow-up	438.490	114.099	5.775	-0.050	509.975	192.151	8.565	-0.108
Sample size: first follow-up	995	995	995	995	423	423	423	423
Sample size: second follow-up	1,151	1,150	1,150	1,152	497	497	497	497
Sample size: third follow-up	1,063	1,063	1,063	1,063	468	469	469	470

Figure 4: Impact on Business Sales and Profits. Source: McKenzie (2017)

	Took a formal loan	Received equity investment	Value of inventories	Made large $K$ purchase	Value of capital purchases	Value of capital stock
<i>Panel A. New firms</i>						
First follow-up	-0.003 (0.006)	-0.005 (0.010)	349 (123)	0.289 (0.031)	1,062 (128)	1,448 (196)
Second follow-up	0.003 (0.009)	0.026 (0.012)	1,869 (350)	0.404 (0.029)	1,543 (143)	4,568 (464)
Third follow-up	0.015 (0.012)	0.001 (0.010)	697 (196)	0.103 (0.031)	155 (122)	3,489 (324)
Control mean: first follow-up	0.011	0.029	721	0.211	345	1,024
Control mean: second follow-up	0.018	0.017	925	0.206	252	1,290
Control mean: third follow-up	0.022	0.020	713	0.206	292	984
Sample size: first follow-up	995	995	991	995	991	995
Sample size: second follow-up	1,071	1,071	1,013	1,071	1,013	956
Sample size: third follow-up	857	857	771	857	771	809
<i>Panel B. Existing firms</i>						
First follow-up	-0.025 (0.017)	0.026 (0.019)	729 (268)	0.369 (0.046)	1,356 (185)	2,050 (335)
Second follow-up	-0.039 (0.020)	0.030 (0.023)	1,320 (579)	0.242 (0.045)	1,018 (202)	3,852 (744)
Third follow-up	0.001 (0.025)	0.001 (0.018)	845 (486)	0.115 (0.052)	221 (340)	4,295 (713)
Control mean: first follow-up	0.042	0.026	1,223	0.358	537	1,759
Control mean: second follow-up	0.063	0.045	2,226	0.434	596	3,190
Control mean: third follow-up	0.061	0.031	1,645	0.362	668	2,536
Sample size: first follow-up	423	423	422	423	423	422
Sample size: second follow-up	458	458	453	458	453	381
Sample size: third follow-up	372	372	360	372	360	331

Figure 5: Impact on Capital. Source: McKenzie (2017)



(Nigeria's currency) higher than the 800,000 naira control mean among new applicants at the time of round 3, and 4.3 million naira higher than the 2.5 million naira control mean among existing applicants.

The winning firms are therefore substantially higher in terms of capital stock, as well as in terms of employment. These results suggest that the main effect of winning is to allow firms to overcome credit constraints by using the capital grants to purchase more capital inputs, hire more labor, and use this to produce a wider variety of inputs. The business plan competition seems an effective tool for identifying entrepreneurs with much greater scope for growth than the typical microenterprise.

## **Validity**

Finally, the author includes a very relevant discussion on the validity of the experiment.

Firstly, a potential concern may be whether the growth of the winners came at the expense of other firms. There are two elements of this concern. The first is a concern about internal validity: if the growth of the winners came at the expense of firms in the control group, the control group would no longer provide a valid counterfactual. The author argues this seems unlikely to be an important concern in this case, since the experimental sample is widely scattered over a country of 170 million people, and is not heavily concentrated in any single industry. As a check, the author shows that there is no heterogeneity in treatment effect, nor difference in control outcomes, with the number of other firms selected as winners in the same state and industry. As a result, the estimates should be internally valid, and are informative for showing the constraints facing individual businesses.

Secondly, it is worth to question the extent to which the results may generalize, that is, the issue of external validity. The program studied in this paper is a nationwide program to a large number of firms in Africa's largest country, and hence should be of intrinsic interest. Nevertheless, there are reasons to believe the findings may generalize beyond Nigeria, given that these types of business plan competitions have become increasingly common, particularly in sub-Saharan Africa. While the amounts offered in Nigeria are high, they are far from unique. The author states that the key difference is the large number of winners and randomized selection provide the ability to evaluate and learn from this competition, which has not been possible in the other cases. It seems likely that these competitions in other countries are also inducing applications from entrepreneurs with high growth potential. Whether such entrepreneurs can be identified in alternative ways to business plan competitions and/or be supported through alternative policy instruments to grants are interesting questions which could be developed in future research.

### 4.3 Difference-in-differences

Difference-in-differences is an analytical quasi-experimental approach that allows to derive causal inference even when randomization is not possible. As previously discussed, we cannot draw causal conclusions by observing simple before-and-after changes in outcomes, neither by comparing outcomes between enrolled and non-enrolled groups.

As the name suggests, the difference-in-differences approach combines these two measurements to compare the before-and-after changes in outcomes for treatment and control groups and estimate the overall impact of the program.

Firstly, the difference-in-differences takes the before-after difference in treatment group's outcomes. In comparing the same group to itself, the first difference controls for factors that are constant over time in that group. Then, to capture time-varying factors, difference-in-differences takes the before-after difference in the control group, which was exposed to the same set of environmental conditions as the treatment group - this is the second difference. Finally, difference-in-differences “cleans” all time-varying factors from the first difference by subtracting the second difference from it. This leaves us with the impact estimation of the intervention.

One advantage of this method is that it can be extended from panel data to pooled cross-section data. This means it is valid even when the units observed in the pre-treatment period are not the same as those observed after treatment, which is very common, for instance, when data is collected through surveys.

Figure 6 displays a schematic representation of the difference-in-differences method described above.

#### Assumptions

Instead of comparing outcomes between the treatment and comparison groups after the intervention, the difference-in-differences methods compares *trends* between the treatment and comparison groups (Gertler et al. (2016)). Although difference-in-differences allows us to tackle differences between the treatment and comparison groups that are constant over time, it will not eliminate the differences between the treatment and comparison groups that change over time. Therefore, the validity of the difference-in-differences approach relies on the *equal trends assumption* - that is, the assumption that outcomes would display equal trends in the absence of treatment. This requires that no time-varying differences exist between the treatment and control groups.<sup>3</sup>

---

<sup>3</sup>Choosing an appropriate control group becomes, therefore, key to validate the results of this method. A solution to obtaining a valid comparison group in the absence of an exogenous shock is to match the treated units to units not necessarily drawn from the same population, but that are similar in a set of observable characteristics. In practice, this is many times achieved by computing **propensity scores** (the probability that units are treated given a set of observables  $\mathbf{x}$ ) and matching units that have similar scores.

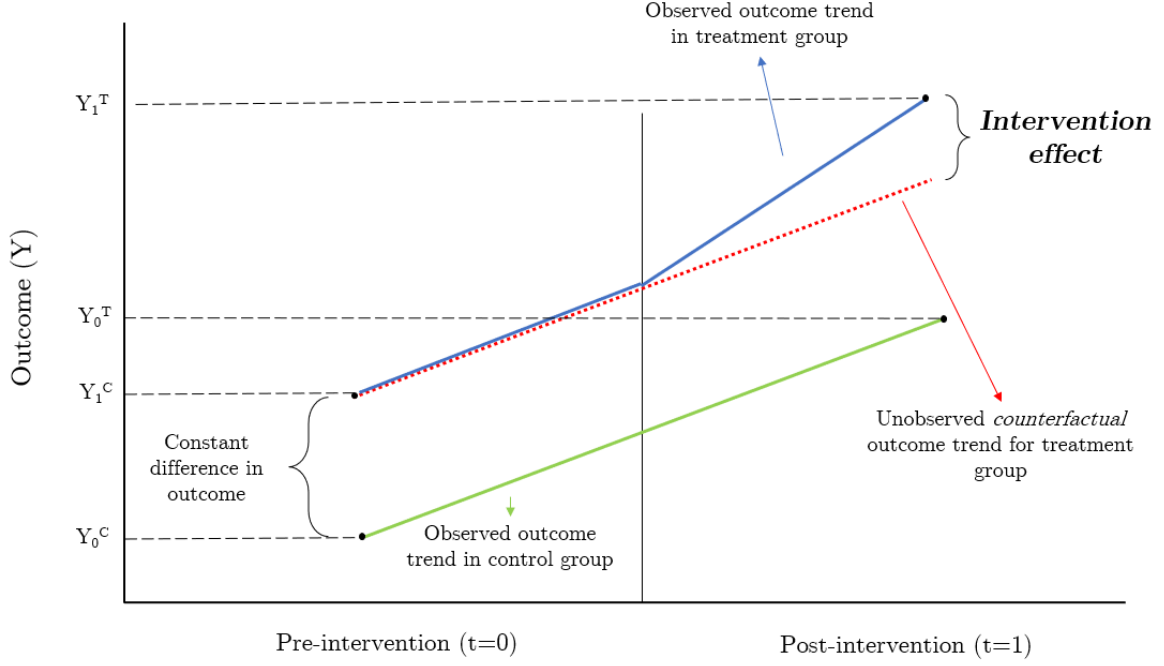


Figure 6: The Difference-in-differences method for causal inference

Even though this cannot be empirically proved (we cannot observe what would have happened to the treatment group in the absence of the treatment, that is, we cannot observe the counterfactual), there are some strategies which can be employed to assess the validity of this crucial assumption.

A first validity check is to compare changes in outcomes for the treatment and comparison groups repeatedly before the program is implemented. If the outcome trend moves in parallel before the program began (that is, parallel pre-trends are established), it is easier to argue that they would have continued moving in tandem in the absence of the program.

A second way to test the assumption of equal trends would be to perform what is known as a *placebo test*. This requires the performance of an additional difference-in-differences estimation using a “fake” treatment group: that is, a group that is known not to be affected in any way by the program (for instance, another set of firms that was not affected by the policy studied). A placebo test that reveals zero impact supports the equal-trend assumption.

A third option would be to perform the placebo test not only with a fake treatment group, but also with a fake outcome. A placebo test that reveals zero impact supports the equal-trend assumption. On the other hand, if the difference-in-differences estimation finds an impact of the intervention on the fake outcome, then the control group must be flawed.

Finally, a fourth way to test the assumption of equal trends would be to perform the

difference-in-differences estimation using different control groups. In this case, similar estimates of the intervention impact confirms the equal-trend assumption.

In the case of cross-section data, an additional assumption is needed: Since it is not possible to eliminate individual fixed effects from repeated observation of the same unit, one needs to ensure that the compositions of the treated and untreated groups is stable before and after treatment.

It is also important to note that even when trends are equal before the beginning of the intervention, bias in the difference-in-differences estimation may still appear and go undetected. That's because the method attributes to the intervention any differences in trends between the treatment and comparison groups that occur from the time intervention begins. If any other factors are present that affect the difference in trends between the two groups and they are not accounted for in multivariate regression, the estimation will be invalid or biased.

## Implementation

A difference-in-differences method application requires data on outcomes in the group that receives the program and the group that does not – both before and after the program.

Using this data, the difference-in-differences impact is computed as follows:

1. Calculate the before-after difference in the outcome (Y) for the treatment group.
2. Calculate the before-after difference in the outcome (Y) for the comparison group.
3. Calculate the difference between the difference in outcomes for the treatment group and the difference for the comparison group.

The table below summarizes these implementation steps.

	Treatment Group	Control Group
Before intervention	$Y_0^T$	$Y_0^C$
After intervention	$Y_1^T$	$Y_1^C$

$$\text{Intervention Effect} = (Y_1^T - Y_0^T) - (Y_1^C - Y_0^C)$$

Figure 7: Impact of the Intervention according to a Difference-in-differences approach

In order to obtain a more rigorous estimation of the difference-in-differences causal impact, we could apply a regression analysis, as seen below.

$$Y_{it} = \beta_0 + \beta_1 Time_t + \beta_2 Group_i + \beta_3 Time_t * Group_i + \beta_4 Covariates + \epsilon_{it}$$

Where  $Time_t$  is a dummy variable that indicates if the observation takes place before or after treatment and  $Group_i$  is a dummy variable that indicates if the observation belongs to treatment or control groups. The estimation through a regression model allows to control for possible factors which could potentially influence the difference in trends between the two groups, ensuring that the estimation results are valid. The difference-in-difference impact estimation is captured by  $\beta_3$ , the interaction term between time and treatment group dummies.

The simple differences-in-differences framework with two time periods and binary treatment presented in this section can easily be extended to the cases where multiple time periods are considered, units are treated in different points in time and/or intensity of treatment differs between units. Furthermore, recent literature has discussed solutions to deal with problems that arise when treatment is staggered or when the parallel-trend assumption does not hold. Useful references to deal with these problems are [De Chaisemartin and d'Haultfoeuille \(2020\)](#), [Goodman-Bacon \(2021\)](#) and [Sun and Abraham \(2021\)](#).

### ***The Effectiveness of Fiscal Incentives for R&D* by Guceri and Liu [Guceri and Liu \(2019\)](#)**

In [Guceri and Liu \(2019\)](#), the authors exploit a policy reform in the UK to derive the impact of tax credits for R&D through a difference-in-differences approach.

The paper starts by highlighting an issue in the literature on tax incentives for R&D, which is a lack of exogenous variation in exposure to the policy. To overcome it, the authors exploit the quasi-experimental setting generated by a UK policy reform in 2008, which saw an increase in the generosity of R&D tax incentives as well as an expansion of the Small and Medium Enterprise (SME) definition, doubling the thresholds measured in indicators such as employment, turnover, and total assets below which a company would be qualified as an SME. As a result, a number of companies that would have been classified as large under the old system became qualified as SMEs and are entitled to more generous deductions.

This constitutes an ideal setting for policy evaluation, since this definition change only applies for the purpose of the R&D tax credit (and no other incentive scheme in the United Kingdom) and there are no concurrent policy changes at the national level that are directly targeted at this group.

By generating differential changes in the user cost of R&D for newly classified SMEs

compared to companies that remained as large facing relatively stable R&D user cost, this policy reform provides an ideal quasi-experimental setting for the identification of the causal effect of R&D tax incentives by addressing the simultaneous determination of R&D spending and its tax price.

The nature of the tax credit schemes applied in the UK allows for the analysis of the effectiveness of the R&D tax incentives in a simpler institutional setting, given that these take the form of enhanced deductions and apply to the total amount of R&D every year for all firms investing in R&D. In this particular policy, combining the effect of both the rate increases and the SME definition change, an SME that was previously labeled as “large” before the reform could deduct, for every £100 of qualifying R&D, £125 against its taxable profit in financial year 2007–2008 and £175 in 2009–2010. Newly qualified SMEs also became eligible to claim cash if they incurred zero or negative taxable profits in the current financial year.

The authors estimate the impact using a difference-in-differences approach, employing a large-scale administrative dataset for the universe of UK corporation tax filings during the 2002–2011 period, which links corporation tax records, qualifying R&D expenditures and financial statements. Additionally, they use both the information present in the tax returns regarding SME or large firm status and employment size from firms’ accounting data to identify treated firms, in order to alleviate measurement errors that may arise from the misclassification of firms.

Therefore, a firm is attributed to the **treatment group** if:

1. Carried out qualifying R&D in at least one of the years before 2008;
2. Carried out qualifying R&D in at least one of the years after 2008;
3. Labeled as “large” in the last of such pre-reform years with positive R&D;
4. Firm size between 250–500 employees in the particular year from bullet 3., which converts the firm to SME according to the new definition (used to refine the treatment group, since the post-reform size may be affected by the reform itself).

The authors employ the following difference-in-differences specification:

$$E[R_{it}|D_{it}, X_{it}] = \exp(\gamma + \delta_D D_i + \delta_i D_i T_t + X'_{it} \beta_x + \Phi_t)$$

where  $R_{it}$  is the level of qualifying R&D spending for company  $i$  in year  $t$  in 2009 prices. The variable  $D_i$  is the treatment dummy and  $T_t$  the time dummy. The coefficient  $\delta_i$  on the interaction term captures the differential change in qualifying R&D spending between pre- and post-2008 periods for the treatment group relative to the control group. Importantly, this parameter can be directly interpreted as the percentage change in the qualifying R&D spending with respect to the tax reform.

Furthermore, time-invariant unobserved firm heterogeneity is captured by the incorporation of additional firm-fixed effects and aggregate macroeconomic shocks that are common to all companies, including the effect of the great recession, are controlled for in all specifications by the set of time fixed effects  $\Phi_t$ . Other non-tax determinants of firm-level R&D spending, including the firm's growth rate of turnover and measures of firm size, are included in the  $X$  vector as additional controls. Standard errors are clustered by firm to correct for over-dispersion.

The baseline results from the estimation of the model above are displayed in Figure 8.

	(1)	(2)	(3)	(4)	(5)	(6)
Treated Firm $\times$ Post-reform	0.308 (0.119)	0.309 (0.118)	0.303 (0.112)	0.302 (0.112)	0.275 (0.117)	0.298 (0.141)
Post-reform	0.080 (0.079)					
Revenue (real, lag) control?	No	No	Yes	Yes	No	Yes
Revenue (real, lag) growth control?	No	No	No	Yes	No	Yes
Revenue $\times$ Post-2008	No	No	No	No	No	Yes
Revenue (real, lag, in log) control?	No	No	No	No	Yes	No
Revenue (real, lag, in log) growth control?	No	No	No	No	Yes	No
Firm fixed effects?	Yes	Yes	Yes	Yes	Yes	Yes
Year fixed effects?	No	Yes	Yes	Yes	Yes	Yes
Observations	3,165	3,165	3,165	3,165	3,159	3,165

Figure 8: Baseline Results from diff-in-diff specification. Source: [Guceri and Liu \(2019\)](#)

In all of the regressions, the difference-in-difference coefficient is significant at the 5 percent significance level, indicating a differential increase in R&D spending for treated firms of at least 30 percent.

However, if we remove the effect of firms' reaction to the early announcement of the policy, by excluding the observations from the year prior to the implementation (2007) the differential increase in R&D spending for treated firms becomes approximately 33 percent.

The authors conclude that there is a 33 percent increase in qualifying R&D spending in response to a 21 percent drop in the tax component of the user cost (for main rate tax payers), translating to an estimate for the elasticity of R&D with respect to its user cost of around  $-1.59$ . For tax payers in the small profits tax rate bracket, the elasticity estimate is  $-2.25$ . These are sizable effects of the policy, which is on the higher end of the estimates found in the literature.

Therefore, these findings suggest that large, consistent programs that support R&D spending in the form of tax incentives are effective in generating additional private R&D.

Finally, the authors emphasize that the key identifying assumption for the difference-in-difference approach is that R&D over time would trend similarly in the treated and control groups in the absence of the policy reform. The policy reform that is used for

identification took place in 2008, which coincides with the Global Financial Crisis. This identifying assumption rules out different time-varying shocks to treated and control firms, and therefore they conduct a series of checks to verify that the size groups of interest followed similar pre-reform trends, and also that different size groups in the medium-large range were not differentially affected by the recession in 2008.

To check this assumption, they perform some of the checks described in the previous section. Resorting to a graphical analysis, there is no particular pattern that suggests violation of the common trends assumption in the levels of R&D prior to the implementation of the policy reform (refer to Figure 9). Furthermore, a discussion on identification and the comparability of treated and control groups, along with placebo tests (namely applying alternative treatment definitions), are explored in the paper.

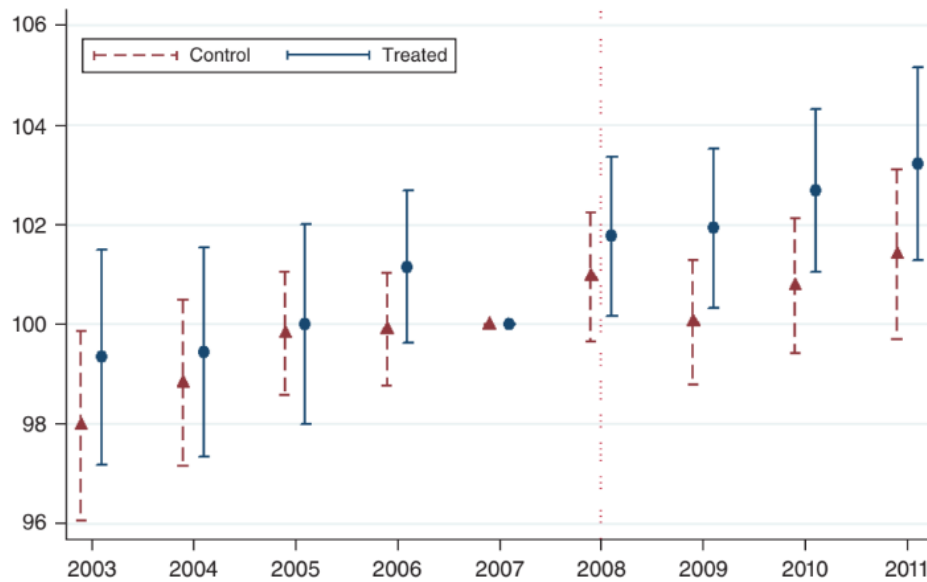


Figure 9: Average R&D Spending across Groups, Relative to 2007 R&D Spending.  
Source: [Guceri and Liu \(2019\)](#)



## 4.4 Instrumental Variables

The quasi-experimental method of instrumental variables (IV) is used to estimate causal relationships when controlled experiments are not feasible or when a treatment is not successfully delivered to every unit in a randomized experiment.

In econometric terms, IVs are used when an explanatory variable of interest is correlated with the error term, in which case it is not possible to establish a causal link between the outcome and the explanatory variable(s) of interest. This is the case when there is reverse causality, that is, the explanatory variables cause the outcome, but the outcome itself influences the explanatory variables, and if there are omitted variables that affect both the dependent and independent variables. In both of these cases, we consider that the explanatory variables are *endogenous*.

To tackle this, we can employ a valid instrument, a variable which induces changes in the explanatory variables but has no independent effect on the dependent variable<sup>4</sup>, allowing a researcher to uncover the causal effect of the explanatory variable on the dependent variable. In a policy evaluation setting, a valid instrumental variable influences the likelihood of participating in a program but does not influence the outcome directly, and must be outside of the participant's control and unrelated to its characteristics.

Figure 10 provides a visual interpretation of the relationship between the explanatory variable ( $X$ ), the outcome ( $Y$ ) and the IV ( $Z$ ).

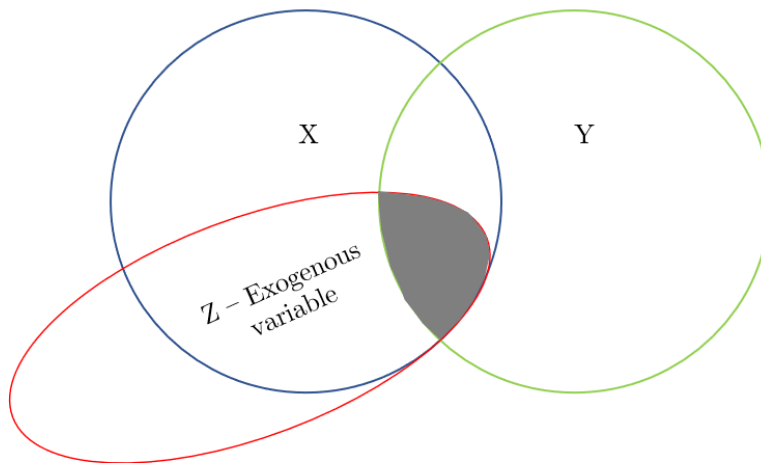


Figure 10: Intuition behind the IV Approach

There are two main requirements for using IVs:

1. The instrument must be correlated with the endogenous explanatory variable(s), conditional on the other covariates. If this correlation is strong, then the instrument is said to have a strong **first stage** or to fulfill the **relevance** criteria. In policy

---

<sup>4</sup>Meaning, it can only affect the independent variable through its effect on the explanatory variable

evaluation, this means that there is a statistically strong relationship between the instrument considered and enrollment in the program. A weak correlation may provide misleading inferences about parameter estimates and standard errors. It is possible to test for this condition through the F-statistic of the first-stage regression, which measures the significance of the IV to explain the endogenous explanatory variable(s).

2. The instrument cannot be correlated with the error term in the main regression, conditional on the other covariates. In other words, the instrument cannot suffer from the same problem as the original predicting variable. If this condition is met, then the instrument is said to satisfy the **exclusion** restriction. This condition is impossible to test, and has to be argued using economic reasoning and evidence.

In the discussion of Randomized Controlled Trials in section 4.2, we assumed units that are assigned to the treatment and comparison groups comply with their assignment. *Full compliance* is more frequently attained in laboratory settings or medical trials, where researchers can carefully ensure that all subjects in the treatment group take a given treatment and that none of the subjects in the comparison group take it. More generally, in that section, we assumed that programs are able to determine who the potential participants are, excluding some and ensuring that others participate, and thus being able to estimate the *average treatment effect* (ATE) for the population.

However, in real-world programs, it might be unrealistic to think that the program administrators will be able to ensure full compliance with the group assignment. To illustrate this, consider a setting where a randomly chosen set of workers is given the possibility of enrolling into on-the-job-training. Even if they are randomly drawn from the population of workers, some treated workers will decide not to enroll for training even if offered the possibility. These units are called non-compliers.

In cases where non-compliance is possible, impact evaluations can estimate the effect of *offering* a program, the **intent-to-treat** (ITT), which is attained by comparing groups to which the program has randomly been offered (in the treatment group) or not (in the comparison group) — regardless of whether or not those in the treatment group actually enroll in the program. The ITT is a weighted average of the outcomes of participants and nonparticipants in the treatment group compared with the average outcome of the comparison group. The ITT is important for those cases in which we are trying to determine the average impact of offering a program, and enrollment in the treatment group is voluntary (Gertler et al., 2016).

A natural application of instrumental variables is to use treatment assignment as an IV to estimate treatment effects. The resulting estimate from employing an instrumental variable is called the **local average treatment effect**. This is because the IV estimates measure the treatment effect on the “compliers” that are induced to participate in the

treatment as a result of the change in the IV. In other words, the LATE is a measure of the effect of treatment on the subset of units that are at the margin of participating in the program, also called *compliers* (Cameron and Trivedi, 2005). The local average treatment effect will presumably differ from the Average Treatment Effect (ATE) - the average effect of treatment on the whole population - if there is non-compliance.<sup>5</sup>

## Implementation

To estimate program impacts under randomized assignment with imperfect compliance, first we must estimate the ITT impact - this is just the difference in the outcome indicator (Y) for the group that is assigned to treatment and the same indicator for the group that is not assigned to treatment. Second, we need to recover the LATE estimate for the group of individuals who comply with their assignment from the ITT estimate.

In statistical terms, the randomized assignment serves as an IV. It is a variable that predicts actual enrollment of units in a program, but is not correlated with other characteristics of the units that may be related to outcomes. While some part of the decision of individuals to enroll in a program cannot be controlled by the program administrators, another part of the decision is under their control. In particular, the part of the decision that can be controlled is the assignment to the treatment and comparison groups. Insofar as assignment to the treatment and comparison groups predicts final enrollment in the program, the randomized assignment can be used as an instrument to predict final enrollment. Having this IV allows us to recover the estimates of the LATE from the estimates of the ITT effect for the units who comply with their assignment type.

More broadly, we can consider the following steps to adopt an IV approach to policy analysis:

1. **Choose the instrumental variable(s):** the choice of the instrument is crucial to ensure the estimation of the causal treatment effect. In general, having detailed information on how the policy was targeted and implemented may reveal sources of exogenous variation that could be used as instrumental variables. Common sources of instruments include policy geographical variation, exogenous shocks affecting the timing of policy implementation and policy eligibility rules.
2. **Estimate the effects:** using two-stage least squares (2SLS) when there is one IV, or, more generally, a GMM estimator.
3. **Check the first stage to assess the relation between the instrument and the treatment:** As previously mentioned, *relevance* assumption, stating that the

---

<sup>5</sup>Intuitively, if one thinks that compliers are more motivated participants, then they may also be more likely to have higher potential outcomes from treatment than those who selected themselves voluntarily out of the program.

instrument should be related with the treatment, is testable by analysing the first stage regression and assessing the strength of the relation between the treatment and the instrumental variables. The usual rule of thumb is that the F-statistic associated with the instrumental variable should be higher than 10 [Stock and Yogo \(2005\)](#). In case of weak instruments the bias of the estimated ATT could be even larger than the one obtained from not taking into account the selection on unobservables at all.

4. **Discuss the exclusion restriction:** even though the assumption that the instrument is not related with the error term is not directly testable, the researcher should discuss extensively why it is believed that the instrumental variable and the error term are not correlated, relying for instance on economic theory and intuition.

### **Do tax credits stimulate R&D spending? The effect of the R&D tax credit in its first decade by [Rao \(2016\)](#)**

This paper evaluates the impact of an R&D tax credit in the U.S. between 1981 and 1991, using an instrumental variable strategy based on tax policy changes. The goal of the paper is to determine the impact of changes in R&D tax credits on firm research intensity - measures as the ratio of R&D expenses to sales. The baseline (second-stage) regression is:

$$\left(\frac{R_{it}}{S_{it}} - \frac{R_{it-1}}{S_{it-1}}\right) = \alpha + \gamma(\rho_{it} - \rho_{it-1}) + \chi_t + \epsilon_{it} \quad (1)$$

where  $\frac{R_{it}}{S_{it}} - \frac{R_{it-1}}{S_{it-1}}$  is the year-on-year change in the ratio of research spending to sales, and  $(\rho_{it} - \rho_{it-1})$  is the change in the actual user cost of R&D capital for firm  $i$ .

The author starts by arguing that using OLS to estimate the impact of tax credits on research intensity suffers from a bias resulting from simultaneity. This issue comes from the fact that R&D tax credits are incremental in their nature, that is, a firm's marginal credit rate is a non-monotonic function of its research expenditures.

Specifically, *firms that fail to exceed their bases receive no subsidy, firms that exceed their bases but do not spend more than twice their bases receive the full statutory subsidy rate and firms that exceed twice their bases receive half the statutory credit rate on their marginal spending* [Rao \(2016\)](#). Therefore, a firm's R&D marginal R&D credit rate and the innovation spending are jointly determined. This implies that the user cost of capital (the explanatory variable of interest, affected by the tax credit) is correlated with unobserved factors ( $\epsilon_{it}$ ) that in turn are likely to be correlated with the outcome of interest. For give an example, if there is a positive shock to R&D spending, then the marginal credit rate could either increase, if the firm is below its base, or decrease if the firm was above its base.

To deal with this endogeneity issue, the authors use an instrumental variable that estimates the change in the user cost of capital ( $\rho_{it} - \rho_{it-1}$ ) using a synthetic change in a firm's marginal user cost ( $\rho_{it}^S - \rho_{it-1}^S$ ) unrelated to current research spending. This change is constructed using the difference in the firm's user cost of capital under the current law and under the previous year's law, calculated based on research spending from two years before. Formally, this corresponds to  $(\rho_{it}^S(R_{it-2}) - \rho_{it-1}^S(R_{it-2}))$ . This instrument captures the change in the user cost of capital that was independent from the firm's investment decisions. Identification is possible due to the several legislative changes that took place in the period of analysis. The source of identification of firm's response to the tax credit is the exogenous change in the effective price of R&D due to changes in the tax rules.

This IV meets the exclusion criterion if the constructed synthetic factor affects R&D spending only through the tax factor, conditional on firm and year fixed effects. As discussed in the paper, this implies that there are no time-varying factors specific to the firm that affect research that are correlated with the variation of legislative changes. In this case, this assumption is plausible due to the non-linearity in the firm-specific tax credit function.

The author also clarifies that the local average treatment effect (LATE) estimated refers to the group of firms whose research budgets are influenced by the marginal tax subsidies. The possibility to generalize the results depends, thus, on whether these firms differ significantly from the universe of firms or not. Figure displays the main table of results from this paper. The estimates imply that a 10% reduction in the user cost of R&D is associated with an increase in the firm's research intensity (measures by the ratio defined above) of about 20%. Furthermore, the results point to an increasing in spending over time, after the average firm overcomes the initial adjustment costs. For smaller or younger firms, however, the effect on long-run spending in research seems to reverse the initial impact. In terms of composition, spending in wages and supplies seems to be the most affected by the change in tax credit. Estimates in a smaller sample suggest that firms increase R&D intensity mainly by increasing qualified research (as opposed to total research).

## 4.5 Regression Discontinuity Design

Regression Discontinuity Design (RDD) is a quasi-experimental design applied to settings where the probability to receive treatment is a discontinuous function of some underlying variable(s). Usually, this arises as a result of institutional rules or policy changes. In RDD, the probability of assignment to treatment is fully determined by a clear cut-off on a continuous underlying function of observable variables. Within a narrow bandwidth around the cut-off, assignment to treatment closely resembles randomization.

	OLS	IV	IV	IV	IV	IV	IV	IV
	Full sample	Full sample	Spline	Trimmed	Industry FE	IV validity	Balanced	Dec. FY
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
$\Delta$ User-cost	-0.039 (0.008)	-0.104 (0.025)	-0.109 (0.040)	-0.117 (0.019)	-0.105 (0.025)	-0.117 (0.016)	-0.082 (0.048)	-0.118 (0.024)
User-cost elasticity	-0.777 (0.157)	-1.980 (0.473)	-2.084 (0.767)	-4.044 (0.330)	-1.998 (0.356)	-2.076 (0.290)	-1.175 (0.695)	-2.248 (0.462)
Prob > F	-	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
Observations	20,883	17,876	17,876	17,433	17,876	16,324	8555	9692
Firms	6338	5715	5715	5619	5715	5391	1711	3160

Notes: All regressions include year fixed effects. All data converted to real dollars using the GDP index. Column (1) presents OLS results while columns (2)–(8) instrument for the endogenous tax tax subsidy using predicted subsidy rates. Column (2) is the baseline IV estimate. Column (3) adds a 5-knot spline in the two-year lag of R&D spending. Column (4) drops the 3% most research intense firms. Column (5) adds industry fixed effects. Column (6) adds a synthetic tax instrument constructed from the four-year lag in R&D spending. Column (7) includes only firms that report in all years while column (8) restricts the sample to firms with December fiscal year ends. Standard errors clustered at the two-digit industry level according to SOI industry codes; these data span 68 industries.

Figure 11: User-cost elasticity of firm R&D intensity. Dependent variable:  $\Delta$  (qualified R&D / sales). Source: [Rao \(2016\)](#)

The canonical example of an RDD is the evaluation of merit-based scholarships on students performance: Suppose that there is a pre-defined cutoff above which students receive a merit scholarship ([Thistlethwaite and Campbell, 1960](#)). While it is not reasonable to assume that students receiving scholarships are similar to those not receiving scholarships (which leads to biased results if one simply compares the outcomes in these two groups), it is possible to argue that the group of students around the cut-off is homogeneous. Therefore, any difference in outcomes between students who receive the scholarship and students who do not receive the scholarship (provided their distance to the threshold is small) can be attributed to the award.

Two conditions are required for a valid RDD:

- A continuous eligibility index (a continuous measure based on observable characteristics - such as age or firm size - that ranks the population of interest)
- A clearly defined threshold above which all individuals are treated and below which no individual is treated.

Furthermore, the validity of this method depends also on two assumptions:

- The eligibility index should not be easily manipulated by the individuals. If, for instance, some individuals strategically keep their income below a certain threshold to receive food stamps, then treatment is no longer random since a subset of them self-selected into treatment.
- Individuals around the cut-off point should have very similar characteristics. In other words, the cut-off in the eligibility index should not be associated with a cut-off in the distribution of characteristics. Although it is possible to test for differences in observed characteristics pre-treatment between treated and untreated individuals, the same should hold for unobserved factors. Usually, for relatively small bandwidths around the threshold, this assumption is plausible.

The simplest case of RDD arises when there is **sharp discontinuity** in the assignment rule. In other words, this means that probability of treatment goes from 0 to 1 at the cutoff. The scholarship example given above is a clear example of a sharp discontinuity design: No student below the defined GPA threshold receives a scholarship and all students above the threshold receive it.

On the other hand, it may happen that the probability of receiving treatment is discontinuous around the cutoff but it does not go from 0 to 1 or vice-versa. In this case, we are in the presence of a **Fuzzy RDD**. For instance, in many countries, the legal age for drinking alcohol is 21. This causes a discontinuity of alcohol consumption around that age and creates two groups that will not differ in anything except for alcohol consumption. Although, some people will still consume alcohol before they reach 21 and some people older than 21 will not consume alcohol, the probability of consuming alcohol has a discontinuity at 21 years-old.

In the Fuzzy RDD setting, the discontinuity in the probability of assignment to treatment acts as an instrumental variable for treatment. Therefore, again, the treatment effect is estimated for the group of compliers.

## Implementation

Implementing an RDD requires the identification of the continuous function that determines assignment to a given treatment and a clear cut-off that changes the probability of this assignment. In practice, for robust estimation, many observations around the cut-off are needed. Furthermore, interval validity relies heavily on the assumption that (1) the behavior of the units of interest does not change due to the existence of the cut-off; and (2) other relevant factors do not change for the same cutoff. Since these two assumptions are not easily testable, the evaluator needs to rely on the analysis of other relevant policies that may influence the outcome of interest and understand to which extent do participants know about the eligibility rule and are able to manipulate their score (Loi and Rodrigues (2012)).

### *Are incentives for R&D effective? Evidence from a Regression Discontinuity Approach* by Bronzini and Iachini (2014)

This paper evaluates the effects of a R&D subsidy program implemented in Italy, using a sharp quasi-experimental design.

The authors aim to examine whether granting subsidies positively impacts the amount of investment in R&D, or, alternatively, if firms would have made had the same amount of R&D had they not received the subsidy. The problem from using a simple

regression of R&D outlays on a binary variable for firms receiving subsidies is that subsidized and non-subsidized firms are likely to differ in terms of unobservable characteristics that are correlated with the amount of R&D spending.

To overcome this endogeneity issue the authors study a unique policy that would select firms to be subsidized based on a score assigned by an independent committee of independent experts. The score is a continuous measure from 0 to 100 that is computed based on the impact of the product on a set of categories. Only projects with a score of 75 points or higher receive the grants.

The authors apply a sharp regression discontinuity design, by comparing the performance of subsidized and non-subsidized firms with scores very close to the 75 points threshold. The application of an RDD to this program meets the two conditions for a valid RDD, since the eligibility index (the score) is continuous and there is a clear threshold that changes the probability of treatment assignment (the 75 points cut-off).

Moreover, the assumptions that guarantee the internal validity of the method hold. First, it is argued that firms cannot manipulate the score perfectly, as it is assigned by an external panel of independent experts. Secondly, the authors test mean differences in key outcomes between treated and non-treated firms around the cutoff and show that they are non-significant in the period before the assignment of the subsidy. This test is displayed in Figure 12.

The authors use both parametric and non-parametric methods to test for the discontinuity at the cut-off point. The baseline regression is:

$$Y_i = \alpha + \beta T_i + (1 - T_i) \sum_{p=1}^3 \gamma_p(S_i)^p + T_i \sum_{p=1}^3 \gamma'_p(S_i)^p + \epsilon_i \quad (2)$$

where  $Y_i$  is the outcome variable;  $T_i = 1$  if firm  $i$  is subsidized (above the threshold of 75 points) and  $T_i = 0$  otherwise;  $S_i = Score_i - 75$ ; the parameters of the score function ( $\gamma_p$  and  $\gamma'_p$ ) are allowed to differ to allow for heterogeneity of the function across the threshold; and  $\epsilon_i$  is the random error. The model is estimated for the whole sample and is later restricted to smaller samples around the cutoff point.

As displayed in Figure 13, the authors find no significant effect on investment in general. Nevertheless, the results do not reject the hypothesis that public financed investment crowded out privately financed investment. Remarkably, they find that when estimating the effect of the program by firm size, there is a positive significant effect for small enterprises. It is argued in the paper that this result is intrinsically related to the fact that adverse selection problems in the financial markets affect negatively access to capital and increase its price for small firms.



Variable	Full sample	50 percent cutoff neighborhood sample (score 52–80)	35 percent cutoff neighborhood sample (score 66–78)
Sales	44,694** (20,442)	4,116 (7,561)	8,179 (10,119)
Value-added	10,070** (4,724)	1,328 (2,057)	1,888 (2,778)
Assets	39,153** (17,576)	5,692 (7,792)	7,792 (10,415)
Return on assets	0.889 (1.179)	0.504 (1.421)	1.415 (1.351)
Own capital/debts	−0.054 (0.082)	−0.212* (0.115)	−0.232 (0.152)
Gross operating margin/sales	0.011 (0.009)	0.001 (0.013)	−0.003 (0.013)
Cash flow/sales	0.019** (0.008)	0.010 (0.011)	0.012 (0.013)
Financial costs/debts	−0.005 (0.004)	−0.006 (0.008)	−0.007 (0.011)
Labor costs/sales	−0.009 (0.010)	0.003 (0.014)	−0.016 (0.019)
Service costs/sales	−0.012 (0.014)	0.015 (0.017)	0.027 (0.021)
Total investment/sales	0.003 (0.009)	0.009 (0.015)	0.024 (0.019)
Tangible investment/sales	0.013 (0.008)	0.020 (0.016)	0.033 (0.020)
Intangible investment/sales	−0.010 (0.006)	−0.011 (0.008)	−0.009 (0.012)

*Notes:* Only manufacturing and construction firms. All the mean-differences refer to the first pre-assignment year (2003 for the first round and 2004 for the second). In the full sample 254 firms are treated, and 103 are untreated. In the 50 percent cutoff neighborhood sample there are 90 treated and 81 untreated firms; in the 35 percent cutoff neighborhood sample there are 57 treated and 58 untreated firms. Investments are calculated as the difference between (tangible and intangible) assets in two consecutive years.

\*\*\*Significant at the 1 percent level.

\*\*Significant at the 5 percent level.

\*Significant at the 10 percent level.

Figure 12: Pre-Assignment Mean-differences between Untreated and Treated Firms.  
Source: [Bronzini and Iachini \(2014\)](#)

Order of polynomial	Total investment/ pre-program sales		Tangible investment/ pre-program sales		Intangible investment/ pre-program sales	
	$\beta$	AIC	$\beta$	AIC	$\beta$	AIC
<i>Panel A. Full sample</i>						
0	0.012 (0.013)	−599.1	0.008 (0.010)	−710.7	0.004 (0.007)	−979.5
1	0.040* (0.020)	−598.8	0.024 (0.015)	−708.6	0.015 (0.012)	−978.5
2	0.045 (0.030)	−595.9	0.021 (0.022)	−704.6	0.024 (0.018)	−978.0
3	0.064 (0.041)	−592.8	0.025 (0.034)	−700.7	0.039 (0.024)	−975.5
<i>Panel B. Local estimates: wide-window sample</i>						
0	0.026 (0.019)	−277.1	0.019 (0.013)	−353.7	0.007 (0.011)	−463.3
1	0.041 (0.034)	−273.8	0.016 (0.022)	−350.0	0.024 (0.020)	−460.8
2	0.110** (0.051)	−274.7	0.036 (0.039)	−347.5	0.073*** (0.024)	−462.6
<i>Panel C. Local estimates: narrow-window sample</i>						
0	0.033 (0.022)	−200.3	0.022 (0.014)	−266.8	0.010 (0.016)	−305.6
1	0.068 (0.040)	−198.8	0.009 (0.034)	−263.5	0.058* (0.027)	−307.1
2	−0.079** (0.035)	−199.8	−0.078 (0.062)	−262.6	−0.000 (0.042)	−305.2
Mean (SD) for untreated firms—full sample	0.033 (0.107)		0.021 (0.084)		0.012 (0.057)	

*Notes:* The table shows the estimates of the coefficient  $\beta$  of model (1) for industrial firms. AIC is the Akaike Information Criterion. Investments are accumulated over the first three years after the assignment (including that of the assignment); sales refer to the pre-assignment year. The polynomial of order 0 is difference in mean between treated and untreated. All the samples have been trimmed according to the fifth and ninety-fifth percentile of the distribution of the Total investment/Pre-program sales ratio (calculated over the full sample). Robust standard errors clustered by score are in parentheses. The number of observations (firms) is 357 in panel A, 171 in panel B, and 115 in panel C.

\*\*\*Significant at the 1 percent level.

\*\*Significant at the 5 percent level.

\*Significant at the 10 percent level.

Figure 13: Baseline Results, Effect of the Program on Investment. Source: [Bronzini and Iachini \(2014\)](#)

## 5 Conclusion

This paper examined how to evaluate the main programs to promote innovation in Portugal. It discussed the rationale and goals of the main types of innovation incentive programs. It showed that existing international evidence on the effects of these programs vis a vis these goals is mixed. Effects on the outcomes of interest are not guaranteed. Success requires rigorous program evaluation and program adjustment in light of the results.

In light of this evidence, the paper presents a toolkit for program evaluation of innovation policies in Portugal. This toolkit has two parts. In the first part, the paper listed the existing databases with relevant information for policy evaluation in the field of innovation. It described how key indicators of firms performance (including sales, size, firm productivity, profitability and scale) can be computed and linked with program support and innovation adoption through common firm identifiers (the fiscal number of the firm).

In the second part, the toolkit focused on the methods that could be employed in future evaluation using that data. These include randomized control trials, differences-in-differences, instrumental variable approach, and regression discontinuity design. It provided step-by-step guidance on how to implement them in Portugal and discussed concrete examples of similar exercises in other countries and their results.

Through this review and discussion, the paper calls for accurate and solid evaluation of innovation incentive programs in Portugal. International evidence provides us with insights regarding similar policies and their outcomes. Yet, much more needs to be done in order to assess the particular effects of Portuguese incentive programs. This is particularly important in the case of the evaluation of the long term persistent effects on firm performance (size and profitability), and dynamics between firms. To assess them, counterfactual analysis (i.e. a comparison with what would have happened without the policy) should be employed. This type of analysis will allow us to answer key policy questions and should be the object of future research.

## References

- Bronzini, Raffaello and Eleonora Iachini**, “Are incentives for R&D effective? Evidence from a regression discontinuity approach,” *American Economic Journal: Economic Policy*, 2014, 6 (4), 100–134. (Cited on page(s) [3](#), [39](#), [41](#), [42](#))
- Cameron, A Colin and Pravin K Trivedi**, *Microeconometrics: methods and applications*, Cambridge university press, 2005. (Cited on page(s) [35](#))
- Chaisemartin, Clément De and Xavier d’Haultfoeuille**, “Two-way fixed effects estimators with heterogeneous treatment effects,” *American Economic Review*, 2020, 110 (9), 2964–96. (Cited on page(s) [29](#))
- Gertler, Paul J, Sebastian Martinez, Patrick Premand, Laura B Rawlings, and Christel MJ Vermeersch**, *Impact evaluation in practice*, World Bank Publications, 2016. (Cited on page(s) [26](#), [34](#))
- Goodman-Bacon, Andrew**, “Difference-in-differences with variation in treatment timing,” *Journal of Econometrics*, 2021, 225 (2), 254–277. (Cited on page(s) [29](#))
- Guceri, Irem and Li Liu**, “Effectiveness of fiscal incentives for R&D: Quasi-experimental evidence,” *American Economic Journal: Economic Policy*, 2019, 11 (1), 266–91. (Cited on page(s) [2](#), [29](#), [31](#), [32](#))
- Jones, Charles I and John C Williams**, “Too much of a good thing? The economics of investment in R&D,” *Journal of economic growth*, 2000, 5 (1), 65–85. (Cited on page(s) [4](#))
- Köhler, Christian, Philippe Laredo, and Christian Rammer**, “The impact and effectiveness of fiscal incentives for R&D,” 2012. (Cited on page(s) [4](#))
- Loi, Massimo and Margarida Rodrigues**, “A note on the impact evaluation of public policies: the counterfactual analysis,” 2012. (Cited on page(s) [39](#))
- McKenzie, David**, “Identifying and spurring high-growth entrepreneurship: Experimental evidence from a business plan competition,” *American Economic Review*, 2017, 107 (8), 2278–2307. (Cited on page(s) [2](#), [19](#), [21](#), [23](#), [24](#))
- Mitchell, Jessica, Giuseppina Testa, Miguel Sanchez Martinez, Paul N Cunningham, and Katarzyna Szkuta**, “Tax incentives for R&D: supporting innovative scale-ups?,” *Research Evaluation*, 2020, 29 (2), 121–134. (Cited on page(s) [4](#))
- Rao, Nirupama**, “Do tax credits stimulate R&D spending? The effect of the R&D tax credit in its first decade,” *Journal of Public Economics*, 2016, 140, 1–12. (Cited on page(s) [3](#), [36](#), [38](#))
- Stock, James and Motohiro Yogo**, “Asymptotic distributions of instrumental variables statistics with many instruments,” *Identification and inference for econometric models: Essays in honor of Thomas Rothenberg*, 2005, 6, 109–120. (Cited on page(s) [36](#))

**Sun, Liyang and Sarah Abraham**, “Estimating dynamic treatment effects in event studies with heterogeneous treatment effects,” *Journal of Econometrics*, 2021, *225* (2), 175–199. (Cited on page(s) [29](#))

**Thistlethwaite, Donald L and Donald T Campbell**, “Regression-discontinuity analysis: An alternative to the ex post facto experiment.,” *Journal of Educational psychology*, 1960, *51* (6), 309. (Cited on page(s) [38](#))

**White, Howard, Shagun Sabarwal, and Thomas de Hoop**, “Randomized controlled trials (RCTs),” *Methodological Briefs, Impact Evaluation*, 2014, (7). (Cited on page(s) [15](#))

**World Bank**, “Randomized Control Trials - Dimewiki.” Accessed: 2022-11-15. (Cited on page(s) [15](#), [16](#))